12-2018

Factors Associated With Participant Attrition in a Longitudinal, Survey-Based Rheumatic Disease Databank

Kathryn A. Grafel
University of Nebraska Medical Center

Recommended Citation

Grafel, Kathryn A., "Factors Associated With Participant Attrition in a Longitudinal, Survey-Based Rheumatic Disease Databank" (2018). *Capstone Experience*. 63.
https://digitalcommons.unmc.edu/coph_slce/63

Factors Associated With Participant Attrition in a Longitudinal, Survey-Based Rheumatic Disease Databank

Kathryn A. Grafel, B.J.

University of Nebraska Medical Center

December 8, 2018

Abstract

Introduction: Participant attrition is a problem common to many longitudinal studies, and when it occurs nonrandomly, it can impact the study's validity and generalizability. Identifying factors associated with attrition can help to detect bias and aid in developing targeted interventions to reduce attrition. **Methods:** Logistic regression and Cox proportional hazards regression models were constructed separately for patients with rheumatoid arthritis, systemic lupus erythematosus, and other rheumatic diseases who participated in the FORWARD study between 1998 and 2018. **Results:** Patient characteristics associated with attrition included male sex, younger age, non-White race, and less education, each of which was identified in multiple models. Score indicating poorer function or greater disease activity on the Health Assessment Questionnaire Disability Index (HAQ), Short Form Health Survey Mental Component Scale (MCS), and Rheumatic Disease Comorbidity Index (RDCI) were associated with dropout in certain groups. Method of recruitment to the study was significant, though its specific impact varied by diagnostic group and analytical technique. **Discussion:** Male sex and less education as predictors of dropout concurred with previous studies, as did the relatively greater significance of socioeconomic factors than health-related factors. The associations with poorer scores on health indices were consistent with researchers' logic that patients in poorer health would be more likely to drop out. **Conclusion:** Patients of male sex, non-White race, younger age, and less education and patients with poor score on health indices were more susceptible to early dropout. FORWARD is advised to develop interventions targeted at retaining at-risk participants, such as achievement tracking to engage younger audiences, accommodations for patients with less education, outreach to patients who report infections, and automated communications triggered by poor scores on health indices.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike information criterion
ARAMIS	Arthritis, Rheumatism, and Aging Medical Information System
BRASS	Brigham and Women's Rheumatoid Arthritis Sequential Study
DMARDs	Disease-modifying anti-rheumatic drugs
HAQ	Health Assessment Questionnaire Disability Index
HIPAA	Health Insurance Portability and Accountability Act
MCS	SF-36 Mental Component Summary
NRAS	National Rheumatoid Arthritis Study
OA	Osteoarthritis
PCS	SF-36 Physical Component Summary
RA	Rheumatoid arthritis
RDCI	Rheumatic Disease Comorbidity Index
RUCA	Rural Urban Commuting Area Codes Data
SF-36	Short Form Health Survey
SLE	Systemic lupus erythematosus

Factors Associated With Participant Attrition
in a Longitudinal, Survey-Based Rheumatic Disease Databank

Introduction

Participant attrition is a common obstacle in longitudinal research. Nonrandom attrition can significantly damage a study's validity and generalizability, particularly in medical research. FORWARD, like many registries and databanks, has endeavored to limit attrition since its inception in 1998. Recognizing factors associated with participants who drop out of a study can help to detect bias and aid in developing targeted interventions to reduce attrition. This study aims to identify both baseline and dynamic factors associated with attrition by use of statistical modeling.

Placement Site

FORWARD, also known as the National Databank for Rheumatic Diseases, is a large database of patient-reported information about rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), osteoarthritis (OA), fibromyalgia, and other rheumatic diseases. For approximately 20 years, FORWARD has been collecting data through surveys that approximately 10,000 patients complete every 6 months. More than 50,000 patients have participated throughout the project's history. FORWARD uses the data to conduct its own research and makes the data available to other researchers (National Data Bank for Rheumatic Diseases, 2017; Wolfe & Michaud, 2011).

Registries and databanks exist for many diseases and groups of diseases. Registries tend to be single-purpose, whereas databanks may encompass multiple diseases and serve many functions (Wolfe & Michaud, 2011). The results of research using information from registries and databanks is often considered more generalizable than that of clinical trials because participants tend to vary more widely than do subjects in clinical trials (Agency for Healthcare Research and Quality, 2014; Iannaccone, et al., 2013; Krishnan, et al., 2004). Whereas subjects in a clinical trial must meet specific criteria (Friedman, Furberg, & DeMets, 2010), generally the only condition for inclusion in a registry or databank is a diagnosis of the disease being studied (Agency for Healthcare Research and Quality, 2014). A further benefit of registries and databanks is their tendency to employ long-term follow-up. This enables collection of useful information about disease progression, which is essential in the case of rheumatic diseases (Krishnan, et al., 2004).

FORWARD participants complete surveys online, on paper, or by phone. Patients who do not wish to complete the 28-page comprehensive survey can opt for a 16-page version. An even shorter questionnaire, only 2 pages long and referred to as the brief version, has also been used in the past. Information is collected on patient demographics, medications, physical function, mental health, and many other areas. The longer surveys contain questions that enable calculations of several health indices developed by researchers with FORWARD and other studies.

FORWARD is one of several registries that collect longitudinal information from patients with rheumatic diseases. Others include the Brigham and Women's Rheumatoid Arthritis Sequential Study (BRASS), the Arthritis, Rheumatism, and Aging Medical Information System (ARAMIS), and the National Rheumatoid Arthritis Study (NRAS). BRASS is an ongoing study limited to patients with RA who are seen at the Brigham and Women's Hospital Arthritis Center. It began in 2003, has enrolled about 1,500 participants to date, and is based on information collected at annual clinic visits (Brigham & Women's Hospital, 2018; Iannaccone, et al., 2013). ARAMIS, which was coordinated by the Stanford Arthritis Center, began in approximately 1976 and operated until the mid-2000s. It employed the semiannual Stanford Health Assessment Questionnaire (HAQ), which is a component of FORWARD surveys. As of the most recent information available, ARAMIS participation numbered about 14,000, including RA patients, OA patients, and other aging individuals (Bruce & Fries, 2005; Krishnan, et al., 2004). NRAS was operated by the University of Connecticut and enrolled 988 patients with rheumatoid arthritis between 1988 to 1997 (Reisine, Fifield, & Winkelman, 2000).

FORWARD has collected data for more years and on more patients than any of these databanks, making it an unparalleled source of data on people living with rheumatic diseases. But in order to remain so, FORWARD must continually recruit new patients and strive to retain existing patients.

The Problem of Participant Attrition

In disease registries, nonrandom participant attrition can have damaging effects both on internal validity and on generalizability. If attrition is random – that is, the group of patients who leave the study resembles the study cohort – then internal validity is maintained. However, if attrition is characteristic of a subset of participants, then the remaining participants no longer resemble the population being studied, thus degrading internal validity. This loss of validity can in turn lead to incorrect conclusions. Medical research, compared to other areas of research, is particularly prone to such bias because patients suffering from higher degrees of disease activity may selectively leave the study. The result is a study sample composed primarily of patients with lower disease activity, which is no longer generalizable to the entire patient population (Barry, 2005; Iannaccone, et al., 2013; Reisine, Fifield, & Winkelman, 2000).

Data from BRASS, ARAMIS, and NRAS has previously been analyzed to identify characteristics of patients who leave longitudinal studies. In each analysis, the authors compared the group of participants who dropped out against the group of those who remained in the study. All three found that patients who left the studies were on average less educated than those who stayed in, and men were more likely than women to leave the study. The authors of the BRASS analysis determined that psychosocial and socioeconomic factors had a greater impact on continued participation than did disease activity. Overall, BRASS and ARAMIS attrition rates were 3.23% and 3.8% per cycle, respectively. NRAS reported that 54% of patients dropped out over the course of a 9-year period, with the rate at its highest early in the study (Iannaccone C. , et al., 2010; Iannaccone C. K., et al., 2013; Krishnan, et al., 2004; Reisine, Fifield, & Winkelman, 2000).

Attrition Among FORWARD Participants

FORWARD began enrolling participants in 1998 and quickly grew until 2003, the year that the Health Information Portability and Accountability Act (HIPAA) was enacted. Prior to this time, rheumatologists' offices were able to provide patient information directly to FORWARD, which could then contact patients for recruitment purposes. This conduit closed when HIPAA placed increased privacy restrictions on patient information. Since then, FORWARD has continued to market itself to potential participants through rheumatologists, but it must rely on the patients to initiate contact. The active participant base decreased from 13,489 patients in 2003 to 9,047 participants in 2016, with cycle-to-cycle retention rates of around 88% between 2014 and 2016 (Hanley, 2016).

Attrition among FORWARD participants has a direct detrimental effect on the research powered by its data and an indirectly detrimental effect on public health. Research using FORWARD data drives treatment decisions, insurance coverage, and management of the many aspects of chronic disease (Michaud, 2016; Wolfe & Michaud, 2009). The contribution that FORWARD makes to improved treatment and management of these diseases is invaluable, and it represents a significant public health impact on the 54.4 million patients in the U.S. alone who suffer from RA, OA, and other rheumatic diseases (Centers for Disease Control and Prevention, 2018).

Benefits of Understanding Factors Associated With Attrition

Among the many benefits of identifying study-specific factors associated with retention is the detection of possible bias. Bias introduced by nonrandom attrition cannot be eliminated; however, recognition of it can mitigate the potential for misinterpretation of results. Where possible, researchers using the data can account for attrition bias in their analyses. When this is not feasible, they can at the very least acknowledge it as a limitation of their findings. Identifying attrition bias, acknowledging it, and evaluating its effect on results is thought by some to be an essential responsibility of registries and databases (Iannaccone, et al., 2013).

A second benefit of identifying factors associated with attrition is the opportunity to tailor retention efforts to participants with a high risk of dropout. Baseline data can be used to do this from the beginning of an individual's participation. Information about dynamic factors associated with attrition can be used to establish triggers for retention interventions later in an individual's time with the study.

Baseline data is collected at the beginning of a patient's participation in the study and can include both patient characteristics and their responses to other items on the initial survey. As discussed above, relevant baseline characteristics that have been identified in previous attrition studies include male sex, younger age, and less education (Iannaccone C. , et al., 2010; Iannaccone C. K., et al., 2013; Krishnan, et al., 2004; Reisine, Fifield, & Winkelman, 2000). Identification of such characteristics enables study administrators to effectively allocate resources and target retention efforts toward participants who are statistically more likely to drop out.

Dynamic factors are those that change over time, such as degree of disease activity, disability level, and employment status. When a dynamic factor is identified as being associated with attrition, changes in a patient's answers over time be interpreted as a warning sign. For example, a registry might identify that participants are more likely to drop out after their survey responses indicate an increase in work days missed due to illness. Using this information, study staff could make individual contact with those patients to encourage continued participation and determine whether accommodations are needed.

FORWARD can benefit from each of these aspects of attrition study results. Information about potential attrition bias would be useful to the many researchers who utilize FORWARD data. Knowledge of baseline characteristics associated with attrition would allow FORWARD to target interventions such as reminder postcards to specific subgroups, allocating financial and staff resources where they will have the greatest effect. Lastly, given information about dynamic factors associated with attrition, FORWARD staff could program its database system to raise an alarm when a patient's survey responses suggest increased risk of dropout.

Goals and Objectives

The goal of this study was to model dropout patterns using both logistic regression and survival analysis techniques for 3 separate groups of patients. Logistic regressions utilized baseline data modeling the log odds of dropout after less than 2 years of participation. The logistic models yielded odds ratios that FORWARD can use to identify patients more likely to drop out early in the study. For the survival analyses, Cox proportional hazards models were constructed using data from each patient's last survey along with dynamic variables representing changes in survey responses between the first and last surveys. The outcome modeled was the mean number of phases that a patient with a given set of characteristics remains in the study, where phases are consecutive 6-month periods. The models produced hazard ratios that FORWARD can use to identify individuals at a risk of earlier dropout than others. Logistic regression and survival models were constructed separately for RA patients, SLE patients, and patients with other rheumatic diseases.

The significance of this attrition study is not limited to FORWARD but extends as well to other rheumatic disease databanks and the general scientific community. This analysis is the first such study utilizing a rheumatic disease database of its size, longevity, and breadth of focus. Previous analyses of attrition trends in rheumatic disease registries and databases were based on smaller samples, a more narrowly defined study population, a shorter study period, or a combination of these limitations. Further, this analysis can serve as a model for any longitudinal study seeking to gain awareness of the factors associated with its own attrition patterns.

Ethics

Approval for this study was sought from the University of Nebraska Medical Center Institutional Review Board, which determined that the research being undertaken was not subject to its oversight.

The existing data will be de-identified by FORWARD prior to my receipt of it, allaying privacy and confidentiality concerns. There are no safety concerns. The researcher has no conflicts of interest.

Methods

This study has 2 aims:

1) Develop logistic regression models to identify baseline factors associated with dropout at less than 2 years of participation and report odds ratios for those factors.
2) Develop Cox proportional hazards models to identify factors associated with time to dropout and report hazard ratios for those factors.

Study Design

This study was a retrospective, secondary analysis of existing data.

Data Source

FORWARD provided deidentified data in SAS Xport Transport File format. Data was extracted from each patient's enrollment record, first survey, and last survey. In order to scale this study to a size manageable within the scope of a master's capstone, only a preselected subset of data elements in FORWARD records was provided. These data elements were selected by the researcher and by FORWARD co-director Kaleb Michaud. Consideration was given to the findings of previous studies, items believed to be clinically relevant, and information that would be useful to FORWARD's recruitment and retention strategies. All available data through June 2018 were analyzed.

Eligibility Criteria and Diagnosis Groups

Inclusion criteria follow:

1) confirmed eligibility for participation in FORWARD;
2) completion of enrollment questionnaire; and
3) completion of at least 1 comprehensive FORWARD survey, either online or on paper, after the initial interview.

Exclusion criteria exist for the logistic regression analyses only. Patients whose first observation occurred in 2016 or later (phases 70 through 74, as described below) were excluded because it was not possible for these patients to have participated for a full 2 years, guaranteeing occurrence of the outcome of interest. No exclusion criteria exist for the survival analyses.

FORWARD's scope includes patients with a wide variety of rheumatic diseases, and it was expected that the factors associated with attrition might vary among diagnosis groups. FORWARD administrators were primarily interested in models assessing dropout patterns for patients with RA and SLE. In order to meet the organization's needs, all analyses were performed separately on 3 groups:

1) patients with a diagnosis of RA, without SLE and with or without or other rheumatic disease comorbidity;
2) patients with a diagnosis of SLE, with or without RA or other rheumatic disease comorbidity; and
3) patients with other rheumatic diseases, without RA or SLE comorbidity.

Measurements

A codebook listing specific variables and their values is found in Appendix A.

Participant Characteristics. Participant characteristics that were evaluated included recruitment method, RA diagnosis, SLE diagnosis, sex, race, education level, body mass index, and date of death.

Recruitment method and date of death were obtained from FORWARD enrollment records. Diagnoses, sex, race, education level, and body mass index were taken from the last survey. Date of death was used only to derive other variables and was not considered as a factor in the models. RA and SLE diagnoses are used to assign participants to analysis groups, and RA diagnosis was considered as a factor in the models for the SLE group.

First and Last Survey Data. The survey response data listed in this section was taken from both the first and last surveys for each patient.

Survey Characteristics. Data elements that characterize the survey itself were phase number, questionnaire type, and date of completion. The phase number is an integer representing the 6-month phase for which the survey was completed; this number was the basis for calculating each participant's duration of participation, as described below. An index of phases and dates is found in Appendix B. Questionnaire type is a 4-digit code that identifies the questionnaire format, length, and other properties, all of which were potential factors in the models.

Participant Characteristics. Participant characteristics taken from first and last surveys were age, self-assessed health status, employment status, annual income, marital status, household size, state of residence, and zip code. Zip code was used only to derive rural vs. urban residence and was not considered as a factor in the models.

Indices. Many of the survey questions are used to calculate indices that are commonly used by researchers to assess a patient's status and degree of function. The HAQ is incorporated in full into the longer surveys, as is another questionnaire called the Short Form Health Survey (SF-36). The survey also includes questions for the rheumatic disease comorbidity index (RDCI), which was developed using FORWARD and U.S. Department of Veterans Affairs data. It is a measure of health problems occurring in conjunction with rheumatic disease, where conditions that have greater impact on rheumatic disease patients are weighted more greatly than are others (England, Sayles, Mikuls, Johnson, & Michaud, 2015). The dataset includes numeric scores for the HAQ Disability Index, the HAQ-II, the SF-36 Physical Component Summary (PCS), the SF-36 Mental Component Summary (MCS), and the RDCI. Scoring of of these indices is summarized in Table 1.

Table 1

Directionality of Health Indices

Index	Range	Direction Indicating Poorer Health or Function
Health Assessment Questionnaire Disability Index (HAQ) Health Assessment Questionnaire II (HAQ-II)	0 – 3	Higher
SF-36 Physical Component Scale (PCS) SF-36 Mental Component Scale (MCS)	0 – 100	Lower
Rheumatic Disease Comorbidity Index (RDCI)	1 – 9	Higher

Note. SF-36 = Short Form Health Survey. Sources: Maska, Anderson, & Michaud, 2011; Utah Department of Health, n.d.; England, Sayles, Mikuls, Johnson, & Michaud, 2015.

Medications. Data elements relating to the patient's medication profile were number of medications taken during the survey phase, number of disease-modifying antirheumatic drugs (DMARDs), number of biologic medications, and use of opioids.

Medical Events. Data elements relating to medical events were occurrence of infection, occurrence of myocardial infarction, presence of heart disease other than myocardial infarction, occurrence of stroke, presence of cancer, and influenza immunization status. Each of these is based on patients' self-reports of experiences during the survey's review period.

Derived Variables. Variables that were assigned or calculated were diagnosis group, questionnaire format, questionnaire length, death phase, censoring status, duration of participation, dropout at less than 2 years, days elapsed before survey completion, longitudinal changes in variables assessed on first and last surveys, and degree of urbanization as defined by the Rural Health Research Center. A detailed explanation of each of these derivations follows.

Diagnosis group was a nominal variable categorized as RA, SLE, and other rheumatic diseases. These were determined using the RA diagnosis and SLE diagnosis variables provided by FORWARD. Assignments were made as described above in the Diagnosis Groups section.

Questionnaire format and questionnaire length were identified within the questionnaire type value. Questionnaire format was categorized as web, print, and telephone. Questionnaire length was categorized as comprehensive, short, and brief. Mapping of these values is detailed in the codebook.

Simplified marital status is a dichotomous recharacterization of the marital status variable into single and partnered categories. Mapping of these values is detailed in the codebook.

Death phase is the survey phase corresponding to date of death. This variable was used only to determine censoring due to death and was not considered as a factor in either model.

Censoring status assigned to participants who were known to have died in the phase immediately following their last survey and to participants who completed phase 74, the final phase for which data was available. Participants were not censored if they did not meet either of those criteria. This variable was used for the survival analysis.

Duration of participation was calculated as the last survey phase minus the first survey phase. The unit of this variable is phases, where 2 phases correspond to a period of 1 year. This variable is the outcome in the survival analysis.

Example: $Phase\ 71\ [July\ 2016] - Phase\ 56\ [January\ 2009] = 15\ phases$

Dropout at less than 2 years was coded as the event of interest when duration was less than 4 phases and non-event when duration was greater than or equal to 4 phases. This variable is used in the logistic regression analysis.

Days elapsed is the number of days in the current phase that elapsed until the patient completed the survey. It was calculated as the SAS date value for the date of survey completion minus the SAS date value for the last day of the previous phase.

Example: $February\ 6,\ 2014 - December\ 31,\ 2013 = 19760 - 19723 = 37$

Longitudinal changes are calculated as the value for the last survey minus the value for the first survey. This calculation is performed for HAQ, HAQ-II, PCS, MCS, RDCI, number of drugs, number of

DMARDs, number of biologics, and health status. These data elements were created only for patients who completed more than one survey, with missing values stored for patients who completed only a single survey.

Degree of urbanization was determined using Rural Urban Commuting Area Codes Data (RUCA), a project of the Rural Health Research Center. RUCA codes , which take discrete, nominal values labeled with numbers from 1.0 to 10.6, were mapped to zip codes in FORWARD data using a table available from the Rural Health Research Center. RUCA categories are groupings of RUCA codes specified as urban, large rural, small rural, and isolated. Mapping of RUCA codes to RUCA categories was performed per Rural Health Research Center definitions (WWAMI Rural Health Research Center, 2006) and is detailed in the codebook. Only the RUCA category variable was used in analyses. RUCA codes and zip codes were used for determining RUCA category but were not considered as factors in the models.

Some categorical variables were collapsed into fewer levels due to small cell counts or highly uneven group sizes or for ease of interpretation, and some continuous variables were categorized to better fit models. Recruitment method was reduced from 34 levels to 4 levels (provider referral, self-enrolled, drug registry, and other). Race was recoded as White, Black, and other. Marital status was dichotomized to single and partnered. Health status was dichotomized to excellent/good and fair/poor. RUCA category was dichotomized to urban and rural/isolated. Details of these recharacterizations are specified in the codebook. Education, originally a continuous variable ranging from 0 to 17, was categorized into 12 years or less, 13 to 16 years, and 17 years or more. Income, originally an ordinal variable with levels ranging from $0 to $150,000, was categorized as less than $30,000, $30,000 to $59,000, and $60,000 or more. Age, originally a continuous variable, was categorized into 4 similarly sized groups: less than 50, 50 to 59, 60 to 69, and 70 or older.

Preliminary Analyses

Prior to beginning statistical analysis, several data preparation steps were followed using SAS software. The original dataset contained one observation for each patient's first survey and a second observation for each patient's last survey, if more than one was completed. Static patient characteristics were included in each observation.

Descriptive statistics were obtained for all participant characteristics and for selected additional variables using the MEANS and FREQ procedures. The purposes of the descriptive analyses included assessment of patient characteristics, assessment of missing data, identification of potential interactions, and identification of predictors with insufficient heterogeneity. Histograms, box plots, scatter plots, and other visual representations were generated for selected variables using the SGPLOT procedure.

Logistic Regression Analyses

The purpose of the logistic regression analyses was to identify baseline factors associated with dropping out before completing 2 years of participation. The outcome variable was Dropout. The LOGISTIC procedure was used with the DESC option to model the probability of an event outcome. Logistic regression analysis was performed separately on each diagnosis group.

Variables Considered as Factors. Because the logistic regression model utilized only baseline factors, only data elements from the enrollment record and the first survey were considered for inclusion in the model.

Predictor Variables With Missing Values or Insufficient Heterogeneity. Variables that were shown in descriptive analyses to have excessive missing values or little heterogeneity were eliminated prior to commencing any model building.

Simple Logistic Regression and Model Assumptions. Prior to evaluating any multiple regression models, simple logistic regression models were constructed using each predictor variable remaining under consideration. The effect of a predictor was considered significant if the *p*-value for the parameter estimate was less than 0.05. Odds ratios were reported for predictors that met the model assumptions and whose effects were statistically significant.

Multiple Logistic Regression – Initial Variable Reduction. Variables with *p*-values less than 0.05 in simple analyses were considered for inclusion in the model. This was a reduction from the planned threshold of 0.05 due to an excessive number of qualifying variables. Continuous variables remaining under consideration were assessed for multicollinearity using PROC CORR. Correlations with $p < 0.05$ and $r > 0.8$ were addressed, and determination of which collinear variables to eliminate were made on a clinical basis. Additional variables were eliminated as necessary using clinical judgment. For example, if both HAQ score and a variable that goes into the HAQ calculation remained under consideration, one was eliminated. If greater than 20 variables remained after these steps, additional variables were eliminated based on clinical relevance and level of significance in simple regression models.

Multiple Logistic Regression – Full Model. All remaining variables were included in the initial multiple logistic regression model. This model, designated as the "full model," additionally contain up to 10 2-way interaction terms that were identified during preliminary analyses. Interactions were included in the model only if the simple effects involved in the interaction were also present. No 3-way or higher interaction terms were considered.

Multiple Logistic Regression – Model Selection. The final model was obtained by manual backward selection. Significance level for variables to remain in the model was 0.05. Classes of categorical variables were retained or dropped as a set, and simple effects were dropped only if they were not components of any interactions remaining in the model.

Multiple Logistic Regression – Final Model. Parameter estimates, *p*-values, odds ratios, and 95% confidence intervals for the odds ratios were reported for the final model.

Survival Analyses

The purpose of the survival analyses was to identify factors associated with duration of participation. More specifically, these analyses determined which factors have a significant effect on the probability of continued participation as a function of time. The primary SAS procedures used were PROC LIFETEST and PROC PHREG. Survival analysis was performed separately on each diagnosis group.

Variables Considered as Factors. The final models are intended for FORWARD administrators to use at any point in a patient's participation and must permit the potential for adjustments in response to survey responses. This means that, unlike the logistic regression models, the survival models were not limited to patient characteristics and observations from the first survey as predictors. In these models, possible predictors included observations from the last survey, along with longitudinal changes in measurements.

For all variables where responses to first and last surveys were provided, both observations were considered. For variables that lent themselves well to calculation of the difference between first and last observations, this difference was also considered as a potential predictor. In cases where at least two forms of a survey response (first, last, and change) were viable candidates for inclusion in a model, only one was selected. When all other factors were essentially equal, preference was generally given to the last survey response or the longitudinal change. This selection was based on log-rank tests of Kaplan-Meier survival estimates.

Variable Characteristics. Categorical variables, variables with insufficient heterogeneity, and variables with many missing values were handled in the same manner as described above for logistic regression analyses.

Kaplan-Meier Estimates. Prior to undertaking any regression analysis, a variety of Kaplan-Meier estimate curves were constructed. The first consisted of a single plot of all data, stratified by diagnosis group. Additional curves were plotted to compare levels of selected variables, including both categorical variables and categorized continuous variables, within each diagnosis group.

Initial Variable Selection. Up to 20 predictor variables were selected for consideration based on preliminary analyses and Kaplan-Meier estimates. The list of potential factors was reduced where necessary. Clinical relevance and findings of prior studies contributed to the list of potential factors.

Proportional Hazards Assumption. The proportional hazards assumption was assessed for each selected predictor variable using Schoenfeld residuals and observed vs. expected plots. For observed vs. expected plots, continuous variables were categorized into 2 or 3 levels. Failure to meet the proportional hazards assumption was defined by a *p*-value less than 0.05 on the Schoenfeld residual correlation test or blatant inconsistency between observed and expected plots. When the two tests disagreed, further methods were employed to investigate whether the assumption was met.

Full Model. All predictors that met the proportional hazards assumption were included in the initial Cox proportional hazards model. Up to 3 predictors that did not meet the assumption were retained for stratification. The resulting model was designated as the “full model.”

Model Selection and Significance Threshold. The final model was obtained by manual backward selection in the same manner as described above for logistic regression analyses. The significance level for variables to remain in the model was 0.05.

Final Model. Parameter estimates, *p*-values, hazard ratios, and 95% confidence intervals for the hazard ratios will be reported for the final model.

Assessment of Predictive Capability

Predictive capability of all models was assessed using Akaike information criterion (AIC). This method was chosen after careful consideration of several validation methods. One possibility was external validation by splitting the sample into a training set and a validation set; this method was rejected with the conviction that the best estimated parameters are based on all available data, not a portion of it. Internal validation options included the bootstrap method and cross-validation by leave-one-out, leave-many-out, or V-fold. Bootstrap was the preferred of these. However, PROC LOGISTIC and PROC PHREG support neither bootstrap nor cross-validation, whereas AIC can be assessed by both SAS procedures. As AIC is an asymptotic equivalent of leave-one-out cross-validation (Shtatland, Kleinman, & Cain, 2004), it was determined to be an acceptable alternative.

Power Analysis

Power analysis for logistic regression models (Aim 1) and survival models (Aim 2) was performed to determine the lowest detectable odds ratios and hazard ratios. Parameters for both analyses included 90% power, 0.05 significance, and a sample size of 5,000. These power analyses were considered conservative since the sample sizes for some diagnosis groups were expected to be larger. For the logistic regression model, lowest detectable odds ratios for a binary predictor variable ranged from 1.023 to 1.702 given varying percentage of sample with X=1, varying response probability, and varying correlation between predictor variables. For the survival model, lowest detectable hazard ratios ranged from 1.003 to 1.098 given an 80% event rate, varying predictor variable standard distribution, and varying predictor correlation. Comprehensive details of both power analyses are given in Appendix C.

Analytical Tools

Statistical analyses were performed and plots were generated using SAS/STAT® software version 9.4 for Windows. Power analysis was performed using PASS 16 Power Analysis and Sample Size Software.

Results

Descriptive Statistics

Data was obtained on 54,027 individuals. This included 35,927 in the RA group, 2,752 in the SLE group, and 15,349 in the other rheumatic diseases group. Table 2 gives demographic information on the study group.

Table 2

Participant Characteristics

Characteristic	All Participants *(n = 54,027)*	RA Group *(n = 35,927)*	SLE Group *(n = 2,752)*	Other Rheumatic Diseases Group *(n = 15,349)*
Age, *mean ± SD*	58 ± 14	59 ± 13	50 ± 13	59 ± 14
Sex [a]				
Female	42,161 (80.6%)	27,706 (78.9%)	2,399 (93.8%)	12,056 (82.3%)
Male	10,149 (19.4%)	7,397 (21.1%)	160 (6.3%)	2,592 (17.7%)
Race [a]				
White	43,858 (88.6%)	30,020 (88.3%)	1,719 (73.7%)	12,119 (91.7%)
Black	2,643 (5.3%)	1,750 (5.2%)	355 (15.2%)	538 (4.1%)
Other	3,029 (6.1%)	2,215 (6.5%)	259 (11.1%)	555 (4.2%)
Education [a]				
12 Years or Less	19,415 (39.6%)	14,408 (42.8%)	647 (27.9%)	4,360 (33.5%)
13 to 16 Years	21,663 (44.2%)	14,309 (42.5%)	1,243 (53.5%)	6,111 (46.9%)
17 Years or More	7,974 (16.3%)	4,979 (14.8%)	432 (18.6%)	2,563 (19.7%)
Marital Status [a, b]				
Single	16,884 (33.4%)	10,687 (32.6%)	1,065 (39.4%)	5,132 (34.3%)
Partnered	33,619 (66.6%)	22,138 (67.4%)	1,636 (60.6%)	9,845 (65.7%)
Employment [a]				
Unemployed	1,849 (3.8%)	1,057 (3.3%)	138 (5.2%)	654 (4.5%)
Paid work	18,352 (37.3%)	11,968 (37.3%)	1,067 (40.5%)	5,317 (36.6%)
Retired	14,242 (28.9%)	9,529 (29.7%)	369 (14.0%)	4,344 (29.9%)
Housework	6,296 (12.8%)	4,145 (12.9%)	311 (11.8%)	1,840 (12.7%)
Student	526 (1.1%)	287 (0.9%)	62 (2.4%)	177 (1.2%)
Disabled	7,983 (16.2%)	5,103 (15.9%)	687 (26.1%)	2,193 (15.1%)
RUCA Category [a]				
Urban	39,878 (74.7%)	26,586 (74.9%)	2,064 (75.8%)	11,228 (74.1%)
Rural or Isolated	13,525 (25.3%)	8,935 (25.2%)	660 (24.2%)	3,930 (25.9%)
Recruitment Category				
Provider Referral	8,416 (15.6%)	6,003 (16.7%)	36 (1.3%)	2,377 (15.5%)
Self-Enrolled	13,290 (24.6%)	6,559 (18.3%)	1,094 (39.8%)	5,637 (36.7%)
Drug Registries	14,297 (26.5%)	12,760 (35.5%)	237 (8.6%)	1,300 (8.5%)
Other	18,024 (33.4%)	10,605 (29.5%)	1,385 (50.3%)	6,034 (39.3%)

Note. All values are assessed as of the first survey. RA = rheumatoid arthritis; SLE = systemic lupus erythematosus; RUCA = Rural Urban Commuting Area Codes.

[a] Data was available for all patients. Percentages are proportions of respondents for whom characteristic was known. [b] Single includes never married, separated, divorced, and widowed. Partnered includes married, widowed/remarried, divorced/remarried, and living together.

The population was comprised predominantly of White females aged in their late 40s to early 70s. Proportion of subjects who were female was approximately 80% in the RA and other rheumatic diseases groups and even higher at 94% in the SLE group. The combined population was 89% White, 5% Black, and 6% of other races. The distribution by race in the RA and other rheumatic diseases group was similar to that of the combined population; however, the SLE group was notably more diverse at 74% White, 15% Black, and 11% of other origins. Very few participants were Asian, Pacific Islander, American Indian, Alaskan Native, and Hispanic origin. Mean age at enrollment for the RA and other rheumatic diseases groups was 59, with the SLE group tending lower at a median age of 50. The overall age range of participants was 7 to 104, with the SLE group limited to 14 to 93.

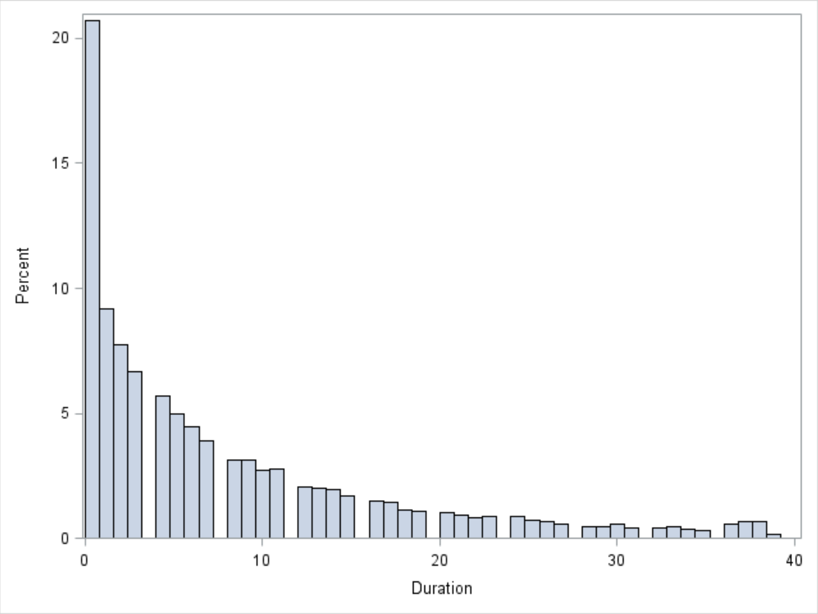
Other common characteristics included urban residence, partnered status, and paid employment or retirement. Across all groups, approximately 75% of patients who provided their location lived in urban areas. Two-thirds of the population who provided details of marital status on their first survey were married or living with a partner. In the RA and other rheumatic diseases groups, approximately one-third of respondents had paid employment and one-third were retired, with 15% to 16% identifying as disabled. In the SLE group, disabled status was notably higher at 26%, a difference that was offset primarily by a lower proportion of retired participants.

The mean education level for the combined population was about 14 years. About 40% of the combined population reported an education level of 12 years or less, about 44% 13 to 16 years, and the remainder 17 or more years. Distribution varied across the diagnosis groups, with the most notable difference being a greater representation of participants with at least some college education in the SLE group. Nearly 9% of patients who provided education information reported having less than a high school education.

Of the 54,027 participants evaluated, more than 20% completed only a single survey. The mean duration of participation was just 8 phases in the RA group and 7 phases in the SLE and other rheumatic diseases group. Forty percent of the RA group dropped out after less than 2 years, along with 44% of the SLE group and 48% of the other rheumatic diseases group. A histogram showing duration of participation for the combined population is shown in Figure 1. Figures showing mean duration by selected patient characteristics are found in Appendix D.

Figure 1

Distribution of Duration of Participation Among All Patients



Certain clinical variables exhibited high rates of missing data due either to having been added to the survey after the study had been underway for several years or to appearing only on the comprehensive version of the survey. Household size, BMI, influenza vaccination status, and side effects were all dropped from consideration in models due to too many missing values in all diagnosis groups. Sufficient information on HAQ and PCS scores were available for first surveys, but they were widely missing in last surveys. Data on MCS scores for first surveys was satisfactory for the SLE and other rheumatic disease groups, but many values were missing in the RA group, and MCS scores for last surveys were lacking across all groups. HAQ-II scores at both first and last observations were sufficient in the SLE group only. Where data was sufficient for first surveys but not for last surveys, variables were retained for logistic regression models but dropped from consideration in survival models. Where data was sufficient in certain diagnosis groups but lacking in others, variables were carried forward on a group-by-group basis.

Information on comorbidities, health status, analgesic use, heart problems, infections, and numbers of drugs, and DMARDs was sufficient for all groups on both first and last observations. Data on biologic drugs, strokes, and cancer was suitable; however, these variables were dropped due to insufficient heterogeneity. Questionnaire format and questionnaire length were dropped due to not all options having been available to participants throughout the study period.

Logistic Regression Models

Odds ratios for dropout within 2 years by diagnosis group are given in Table 3. Odds ratios were calculated from final logistic regression models that were obtained by reduction of full models containing predictors significant in univariable models, along with selected interactions. Detailed results of univariable and multivariable models are found in Appendices E, F, and G.

Table 3

Odds Ratios for Dropout Prior to 2 Years in Final Logistic Regression Models

	OR [95% CI]		
Parameter	**RA Group**	**SLE Group**	**Other Rheumatic Diseases Group**
Sex			
Male vs. Female			1.27 [1.14, 1.42]
Age (+ 10 years)		0.89 [0.85 – 0.96]	
Disabled Patients			0.85 [0.76, 0.95] [a]
Patients Doing Housework			0.88 [0.81, 0.96] [a]
Patients With Paid Employment			0.84 [0.79, 0.90] [a]
Retired Patients			1.11 [1.01, 1.23] [a]
Students			0.74 [0.57, 0.97] [a]
Unemployed Patients			0.83 [0.70, 0.99] [a]
Race [b]			
Black vs. White	1.40 [1.24 - 1.58]		1.31 [1.06, 1.62]
Other vs. White	1.47 [1.32 - 1.63]		1.24 [1.01, 1.51]
Marital Status			
Single vs. Partnered	0.87 [0.82 - 0.92]		0.91 [0.83, 1.00]
Education Level (+ 1 year)	0.95 [0.94 – 0.96]		0.95 [0.93, 0.97]

Table 3

Odds Ratios for Dropout Prior to 2 Years in Final Logistic Regression Models

	OR [95% CI]		
Parameter	**RA Group**	**SLE Group**	**Other Rheumatic Diseases Group**
Recruitment [b]			
Provider Referral vs. Other	0.77 [0.71 - 0.83]	0.82 [0.63 – 1.05] [c]	1.13 [0.96, 1.33] [b, c]
Self-Enrolled vs. Other	1.02 [0.93 – 1.12] [c]	0.65 [0.50 - 0.86]	0.98 [0.83, 1.17] [b, c]
Drug Registries vs. Other	0.80 [0.74 - 0.87]	0.60 [0.42 - 0.87]	1.35 [1.10, 1.65]
Employment [b]			
Housework vs. Disabled	0.97 [0.87 – 1.07] [c]	0.79 [0.58 – 1.07] [c]	0.95 [0.80, 1.13] [a, c]
Paid Work vs. Disabled	1.15 [1.05 - 1.26]	0.70 [0.55 - 0.89]	0.96 [0.82, 1.12] [a, c]
Retired vs. Disabled	1.02 [0.93 – 1.11] [c]	0.69 [0.51 - 0.95]	0.82 [0.68, 1.00] [a]
Student vs. Disabled	1.84 [1.40 - 2.42]	0.71 [0.40 – 1.29] [c]	0.86 [0.38, 1.92] [a, c]
Unemployed vs. Disabled	1.37 [1.17 - 1.60]	1.07 [0.72 – 1.59] [c]	1.17 [0.90, 1.52] [a, c]
Income [b]			
$30,000 - $59,999 vs. Less Than $30,000		1.00 [0.80 – 1.24] [c]	
$60,000 or More vs. Less Than $30,000		0.76 [0.62 - 0.94]	
RUCA Category			
Urban vs. Rural	0.90 [0.84 - 0.95]		
HAQ Score [d] (+ 1 unit)			
Patients Assessing Health as Fair/Poor	0.74 [0.65 - 0.85]		
Patients Assessing Health as Excellent/Good	0.83 [0.72 – 0.97]		
MCS Score (+ 10 units)			0.89 [0.86, 0.93]
RDCI (+ 1 unit)	1.06 [1.04 - 1.07]		
Infection			
(Yes vs. No)	1.12 [1.06 - 1.19]		1.19 [1.09, 1.30]
Self-Assessed Health Status			
Excellent/Good vs. Fair/Poor	1.09 [0.98 - 1.22] [c]		
Analgesic Use			
Yes vs. No			

Note. All values are assessed as of first survey. RA = rheumatoid arthritis; SLE = systemic lupus erythematosus; RUCA = Rural Urban Commuting Area Codes; HAQ = health assessment questionnaire; MCS = mental component score; RDCI = rheumatic disease comorbidity index.

[a] The model for the other rheumatic diseases group contains an interaction between age and employment. Odds ratios given for employment categories are at age 59, the mean age for this group. [b] Selected pairwise odds ratios between levels of these variables are given in Tables E4, F4, and G4. [c] Confidence interval is inclusive of 1. Odds ratio may be informative but should not be considered definitive. [d] The model for the RA group contains an interaction term between HAQ score and self-assessed health status. No odds ratio is given for patients in fair or poor health vs. patients in good or excellent health because the main effect of health status was not significant.

Demographic characteristics including sex, age, race, marital status, and education level were each significant in at least 1 diagnosis group. Males in the other rheumatic diseases group had 27%

greater odds of dropout within 2 years than did female patients. , and non-White patients had 24% to 47% greater odds than White patients in the RA and other rheumatic diseases groups. The effect of age was significant in the SLE and other rheumatic disease groups. In the SLE group, a 10-year age increase corresponded to an 11% reduction in odds of dropout. In the other rheumatic diseases group, age was dependent on employment category, older age being associated with higher dropout odds in retired patients and lower odds in all others. Marital status was significant in the RA and other rheumatic disease groups, with partnered patients having reduced dropout odds of 13% and 15%, respective of the 2 groups. Having 1 more year of education corresponded to a 5% reduction in dropout odds in the RA and other rheumatic disease groups, and residing in an urban area corresponded to a 10% reduction over rural residency in the RA group.

With regard to employment category, the greatest differences in odds were seen in the SLE group. In this group, the odds of dropout for disabled and unemployed patients were approximately 1.5 times those for working or retired patients (odds ratios and confidence intervals given in Table F4). In the RA group, working patients had higher odds of dropout than disabled or retired patients (OR 1.15 and 1.13, respectively). Unemployed patients had greater odds of dropout than working or retired patients (OR 1.19 and 1.35, respectively). Income was a factor in the SLE group, where the odds of dropout for patients in the lower income brackets (less than $30,000 and $30,000 - $59,999) were 1.34 and 1.31 times the odds for patients in the highest bracket ($60,000 or more).

The recruitment predictor had a significant overall effect on the outcome in all models. The original recruitment variable, which had had many levels, was collapsed into 4 categories: provider referral, self-enrolled, drug registries, and other. Although “other” was selected as the reference group due to having the most observations, comparisons between the 3 named categories are of greater interest. Odds ratios for these comparisons are given in Tables E4, F4, and G4. Results varied considerably between diagnosis groups. The most notable odds ratios were as follows:

- self-enrolled vs. provider referral in the RA group: 1.33 [95% CI 1.23, 1.44];
- self-enrolled vs. drug registry RA patients: 1.28 [95% CI 1.18, 1.39];
- practice-enrolled vs. self-enrolled patients in the SLE group: 1.25 [95% CI 1.03, 1.52].
- self-enrolled vs. drug registry patients in the other rheumatic diseases group: 1.28 [95% CI 1.11, 1.48]; and
- drug registry vs. self-enrolled patients in the other rheumatic diseases group: 1.37 [95% CI 1.18, 1.60].

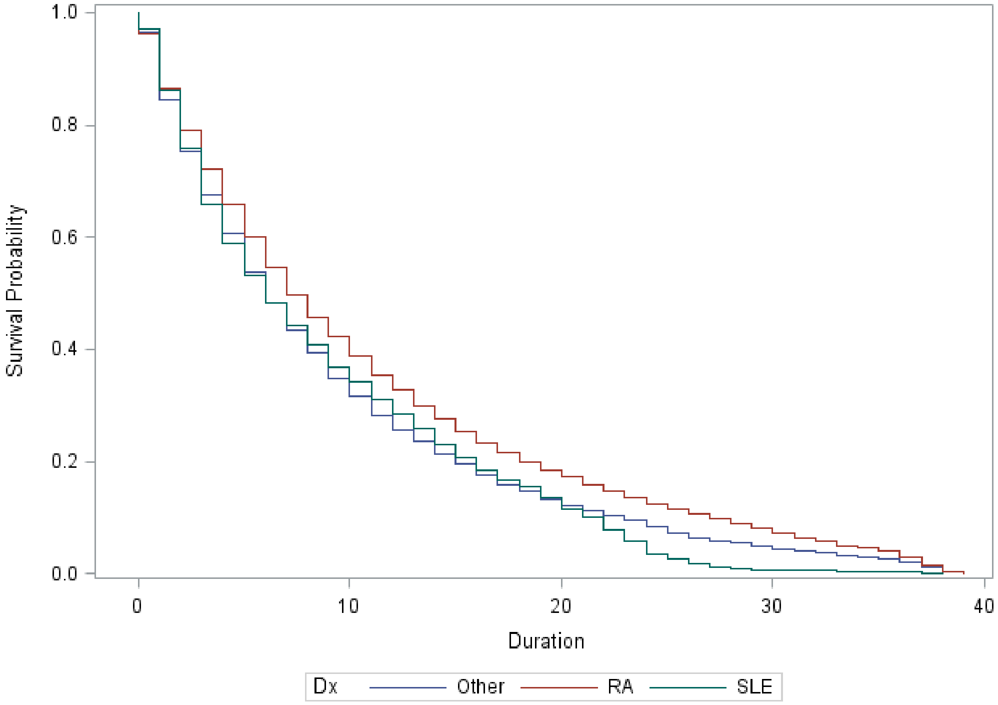
Indicators of health were significant in the RA and other rheumatic disease groups. The most noteworthy effect was that of HAQ score. In RA patients who assessed their health as excellent or good, a 1-unit higher HAQ score (on a scale of 0 to 3) corresponded to a 17% reduction in odds of early dropout. In RA patients patient who assessed their health as fair or poor, the reduction was 26%. In the other rheumatic diseases group, a 10-unit higher MCS score (on a scale of 100 points) was associated with an 11% reduction in odds. Occurrence of infection multiplied the odds by 1.12 and 1.17 in these groups. RDCI had a slight effect in the RA group, where a 1-unit higher score (indicating more comorbidities) corresponded to 1.06 times the odds of dropout.

Survival Models

Kaplan-Meier survival curves for each diagnosis group (Figure 2) showed clear differences in survival probability patterns between patients with RA, patients with SLE, and patients with other rheumatic diseases. The RA group tended toward higher survival probability than the other two groups at all durations greater than 2 phases. The curves for the SLE and other rheumatic diseases groups were similar up to approximately 22 phases, beyond which the other rheumatic diseases group exhibited a notably higher survival probability for the remainder of the duration range. The log-rank test indicated statistically significant differences in the survival functions ($p < 0.001$). Kaplan-Meier survival plots for a variety of predictors within each diagnosis group are found in Appendices H, I, and J.

Figure 2

Kaplan-Meier Survival Curves for RA, SLE, and Other Rheumatic Diseases Groups



Hazard ratios based on multivariable Cox proportional hazards regression models for each group are shown in Table 4. As with the survival analysis, these final models were obtained by backward selection from models containing qualifying predictors. Predictors were selected for the full models based on log rank tests of Kaplan-Meier curves and evaluation of the proportional hazards assumption. Among the selected predictors were a variety of patient characteristics, observations from last surveys, and longitudinal changes from first to last surveys. Detailed results of all models are found in Appendices H, I, and J.

Table 4

Hazard Ratios for Final Survival Models

Parameter	RA Group	SLE Group	Other Rheumatic Diseases Group
Sex			
Male vs. Female	1.16 [1.12, 1.21]		1.11 [1.05, 1.18]
Age (+ 10 years) [a]	0.77 [0.76, 0.78]	0.79 [0.75, 0.82]	0.77 [0.75, 0.78]
Race			
Black vs. White	1.14 [1.06, 1.23]		
Other vs. White	1.03 [0.96, 1.09] [b]		
Marital Status			
Single vs. Partnered	0.95 [0.92, 0.98]		
Education Level (+ 1 year)	0.96 [0.95, 0.96]	0.96 [0.94, 0.98]	0.98 [0.97, 0.99]
Employment [c]			
Housework vs. Disabled	1.22 [1.15, 1.29]		
Paid Work vs. Disabled	1.12 [1.07, 1.18]		
Retired vs. Disabled	1.10 [1.05, 1.16]		
Student vs. Disabled	1.25 [1.01, 1.53]		
Unemployed vs. Disabled	1.26 [1.15, 1.39]		
Change in RDCI (+ 1 unit difference)	0.96 [0.95, 0.97]		0.99 [0.97, 1.00]
Number of Drugs (+ 1 drug)	0.99 [0.98, 0.99]		0.99 [0.98, 0.99]
Health Status			
Excellent/Good vs. Fair/Poor	1.05 [1.02, 1.08]		
Infection			
Yes vs. No		1.20 [1.08, 1.34]	
Change in HAQ II Score (+ 1 unit)		0.85 [0.77, 0.94]	

Note. All dynamic variables are assessed as of the last survey. RA = rheumatoid arthritis; SLE = systemic lupus erythematosus; RDCI = rheumatic disease comorbidity index.
[a] Odds ratio for a 1-year age increase in both the RA group and the other rheumatic diseases group was 0.97 [0.97 – 0.98]. Odds ratio for 1-year increase in the SLE group was 0.98 [0.97 – 0.98]. [b] Confidence interval is inclusive of 1. Odds ratio may be informative but should not be considered definitive. [c] Pairwise odds ratios for selected employment types are given in Table H4.

The most noteworthy predictor in all models was age, with a 10-year age increase corresponding to a 21% to 23% reduction in hazard of dropout. Sex was also an important predictor with hazard ratios for males vs. females of 1.16 in the RA group and 1.11 in the other rheumatic diseases group. Race was significant only for the comparison between Black and White patients in the RA group, where the hazard for Black patients was 1.14 times that of White patients.

Employment was significant in the RA models, where paid work, housework, retirement, and unemployment were all associated with higher hazards than that of disabled status. Hazard ratios for several pairwise comparisons among employment categories are presented in Table G4. The most notable were unemployed patients and patients whose primary occupation was housework at 1.26 and 1.22 times, respectively, the hazard of disabled patients. The hazard ratio for paid work vs. disabled was 1.12, and the ratio for retired vs. disabled was 1.10.

Marital status was significant in the RA group, as was education in all groups, with partnered status and more education having slightly lower hazards.

Among predictors describing patients' health, the most notable effects were occurrence of infection and change in HAQ II Score among SLE patients. Infection increased the hazard of dropout over time by a factor of 1.20. A larger positive change in the HAQ II score was associated with a lower hazard of dropout; reversing the numbers gives an odds ratio of 1.18 for each additional unit lower that the HAQ II score dropped between the first and last surveys. Number of drugs and change in RDCI had small-scale effects in the RA and other rheumatic diseases groups, with a higher number of drugs and worsening comorbidity corresponding to a lower dropout hazard over time. Additionally, excellent or good health carried 1.05 times the hazard of fair or poor health in the RA group.

Discussion/Recommendations

Based on previous research using BRASS, ARAMIS, and NRAS data, expectations prior to this analysis were that male patients and those with less education would be at greater risk of dropout. Both effects were confirmed. The final models demonstrated that male patients in the other rheumatic diseases group had higher odds of dropout within 2 years, and males in both the RA and other rheumatic diseases groups had higher hazard of dropout over time. Although sex was not retained a predictor in the other multivariable models, simple logistic regression and Kaplan-Meier survival plots indicated that males tended toward greater rates of dropout in all groups. Education level was a predictor in all final models except the logistic regression for the SLE group, and in each case less education was associated with higher odds or hazard of dropout. The full logistic regression model for the SLE group did contain education, and the result was the same.

In accordance with Iannaccone's findings in the BRASS study (2013), socioeconomic factors in general were expected to be relevant to dropout patterns. This held true, with the logistic regression models for the RA and other rheumatic diseases groups revealing non-White patients to have 24 to 47% greater odds of dropout within 2 years than White patients. This finding was consistent in the survival model for the same group, though on a smaller scale. Univariable logistic regression for the other rheumatic diseases group indicated that Black patients had 38% greater odds of dropout than White patients, and patients of other races had 33% greater odds than White patients. Marital status was significant in some models, in each case indicating that single patients were more likely to drop out.

In the area of disease and health characteristics, FORWARD administrators expected that greater disease activity, as indicated by scores on functional indices, would be associated with dropout. Further, it was anticipated that worsening scores would precipitate dropout. This was confirmed in the logistic regression model for the RA group, where lower HAQ score and higher RDCI carried greater odds of dropout. Change in HAQ II score was a predictor in the survival model for the SLE group, where a greater mean decrease in score between first and last surveys was associated with greater hazard of dropout. Occurrence of infection was associated with greater risk of dropout in the models in which it appeared, as was also expected.

In general, worse health index scores were an indicator of greater risk of dropout across models. Further analysis is warranted to determine specific thresholds associated with concerning increases in risk. This information would be valuable to FORWARD administrators, who could implement interventions based on specific values.

Impact of Findings

Given the awareness of these factors associated with attrition, FORWARD is empowered to critically evaluate its approaches to retention efforts with a focus on the patients most at risk for early dropout. Just a few of the many opportunities follow:

- development of programs that are likely to be effective with men and younger patients;
- assessment of existing communications and the participant survey itself to ensure that patients with less education are being adequately accommodated;
- allocation of funds to recruitment efforts that result in a more retainable patient base;
- implementation of automated alerts that trigger when a patient's HAQ score declines or RDCI rises or when a patient reports having had an infection in the most recent survey phase.

Beyond FORWARD, other databanks and registries with similar structure might apply the analytical techniques used here to their own data. Knowledge of study-specific factors associated with participant attrition could enable administrators of these studies to improve their own retention programs.

Limitations and Opportunities for Further Research

This study has several limitations, chief among them the narrow selection of data elements from among the many available in FORWARD data, which was necessitated by the scope of this project. Factors may exist that possess greater predictive ability than those considered, whether currently in FORWARD records or in questions not yet asked of patients. Additionally, meaningful interactions may be present that were not tested in models, particularly in the survival models where interactions were not considered. Further regarding statistical procedures, the models presented here did not provide for the use of repeated measures data. Assessment of a dataset containing the full series of each patient's survey responses might reveal patterns beyond those detectable by this limited analysis. Expanding the outcome to account for competing risks, such as the various reasons given for dropout in exit interviews, might also be informative. Added value might also be identified by reducing the other rheumatic diseases group to specific diagnosis groups or by isolating the group of crossover patients with both RA and SLE diagnoses..

Conclusions

The results of this study suggest that patients of non-White race, younger age, and less education are more susceptible to early dropout from FORWARD participation. Clinical factors may also play a part, though their role seems less straightforward. Nevertheless, knowledge of these clinical factors as well as the socioeconomic ones can enable FORWARD to prioritize retention efforts that are tailored toward patients at greater risk of dropout.

FORWARD administrators are advised to take the following actions:

- Pursue the existing proposal for creating an achievement tracking system analogous to those found in video games. This should be featured in a mobile application rather than on the organization's web site. Such a system would likely fare well with younger patients, particularly males.
- Investigate the reasons for patients with less education dropping out of the study at greater rates. Identify other characteristics that are associated with less education and with dropout, then review existing programs with these in mind. Conduct a root cause analysis to seek out the ultimate causes of dropout. If patients with lower education are intimidated by the comprehensive survey, consider inviting them to instead use the short survey or to participate by phone. Consider similar solutions if socioeconomic factors such as shift work, non-partnered status, or larger households leave the participants little time to complete surveys.

- Assess the motivations of self-enrolled patients, who tend toward greater risk of dropout. Find out why they enrolled in FORWARD, what their expectations were, and how the study met or did not meet those expectations. Use this information to design interventions appropriate to those findings, and target such programs to self-enrolled patients.
- Expand on the actions taken by FORWARD staff when patients report infections. When staff contact patients to obtain details of the infections, have them also ask whether accommodations are needed for the following survey. If patients fail to participate in the following phase, employ a simple targeted intervention such as a phone call or a letter wishing them well after their illness and encouraging them to resume participation with the next phase.
- Revisit the data on health index scores and dropout. Identify the HAQ, MCS, and RDCI thresholds that correspond to notable increases in risk, considering that different thresholds may apply to patients with different characteristics. Develop a simple SQL query to identify respondents whose scores are below the determined thresholds, outputting a mailing list for the selected patients. Craft a simple message that encourages continued participation to be sent by e-mail or letter. Automate this process to run monthly against surveys received since the previous run.

References

Agency for Healthcare Research and Quality. (2014). Registry design. In R. E. Gliklich, N. A. Dreyer, & M. B. Leavy (Eds.), *Registries for evaluating patient outcomes: A user's guide* (3 ed.). Agency for Healthcare Research and Quality. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK208632/

Arthritis Foundation. (n.d.). *Arthritis by the numbers: Book of trusted facts & figures.* Atlanta: Arthritis Foundation. Retrieved March 20, 2017, from http://www.arthritis.org/Documents/Sections/About-Arthritis/arthritis-facts-stats-figures.pdf

Barry, A. E. (2005). How attrition impacts the internal and external validity of longitudinal research. *Journal of School Health, 75*(7).

Brigham & Women's Hospital. (2018). *About BRASS*. Retrieved September 24, 2018, from BRASS: http://www.brassstudy.org/brass-information/overview/

Bruce, B., & Fries, J. F. (2005). The Arthritis, Rheumatism and Aging Medical Information System (ARAMIS) - Still young at 30 years. *Clinical and Experimental Rheumatology, 23*(5), S-163-S-167.

Centers for Disease Control and Prevention. (2015, October 28). *Cost statistics | data and statistics | arthritis | CDC*. Retrieved March 20, 2017, from Centers for Disease Control and Prevention: https://www.cdc.gov/arthritis/data_statistics/cost.htm

Centers for Disease Control and Prevention. (2018, July 18). *National statistics | data and statistics | arthritis | CDC*. Retrieved September 24, 2018, from Centers for Disease Control and Prevention: https://www.cdc.gov/arthritis/data_statistics/national-statistics.html

England, B. R., Sayles, H., Mikuls, T. R., Johnson, D., & Michaud, K. (2015, May). Validation of the rheumatic disease comorbidity index. *Arthritis Care & Research, 67*(6), 865-872.

Friedman, L. M., Furberg, C. D., & DeMets, D. L. (2010). Study population. In L. M. Friedman, C. D. Furberg, & D. L. DeMets, *Fundamentals of clinical trials* (4 ed.). New York: Springer.

Gabriel, S. E., Crowson, C. S., Campion, M. E., & O'Fallon, W. M. (1997). Indirect and nonmedical costs among people with rheumatoid arthritis and osteoarthritis compared with nonarthritic controls. *The Journal of Rheumatology, 24*(1), 43-48.

Hanley, J. (2016). National data bank for rheumatic diseases: Recruitment and retention study. University of Nebraska Omaha. Retrieved from Unpublished report

Iannaccone, C. K., Fossel, A., Tsao, H., Cui, J., Weinblatt, M., & Shadick, N. (2013). Factors associated with attrition in a longitudinal rheumatoid arthritis registry. *Arthritis Care & Research*, 1183-1189. doi:10.1002/acr.21940

Iannaccone, C., Fossel, A., Tsao, H., Cui, J., Weinblatt, M., & Shadick, N. (2010). *Factors associated with attrition in a longitudinal rheumatoid arthritis registry [research poster].* Retrieved from https://www.brassstudy.org/wp-content/uploads/2010/08/Factors-of-Attrition-in-a-longitiudinal-RA-registry1.pdf

Krishnan, E., Murtagh, K., Bruce, B., Cline, D., Singh, G., & Fries, J. F. (2004). Attrition bias in rheumatoid arthritis databanks: A case study of 6346 patients in 11 databanks and 65,649 administrations of the Health Assessment Questionnaire. *The Journal of Rheumatology, 31*(7), 1320-1326.

Michaud, K. (2016, January). Notes from the directors: A story about being heard. *The NDB Research Newsletter*. Retrieved March 21, 2017, from https://www.arthritis-research.org/sites/default/files/NDB%20Jan%202016.pdf

Myaasoedova, E., Crawson, C. S., Kremers, H. M., Therneau, T. M., & Gabriel, S. E. (2010). Is the incidence of rheumatoid arthritis rising?: Results from Olmsted County, Minnesota, 1955-2007. *Arthritis & Rheumatology, 62*(6), 1576-1582. doi:10.1002/art.27425

National Data Bank for Rheumatic Diseases. (2017). *About the NDB*. Retrieved March 21, 2017, from National Data Bank for Rheumatic Diseases: https://www.arthritis-research.org/about/about-ndb

Reisine, S., Fifield, J., & Winkelman, D. K. (2000). Characteristics of rheumatoid arthritis patients: Who participates in long-term research and who drops out? *Arthritis Care & Research, 13*(1), 3-10.

Schappert, S. M., & Rechtsteiner, E. A. (2011). *Ambulatory medical care utilization estimates for 2007.* National Center for Health Statistics.

Shtatland, E. S., Kleinman, K., & Cain, E. M. (2004). A new strategy of model building in PROC LOGISTIC with automatic variable selection, validation, shrinkage and model averaging. Cary, NC: SAS Institute Inc.

Utah Department of Health. (n.d.). *Interpreting the SF-12: Comparing versions 1 and 2 of the SF-12.* Retrieved December1 2018, from http://health.utah.gov/opha/publications/2001hss/sf12/SF12_Interpreting.pdf

Wolfe, F., & Michaud, K. (2009). Predicting depression in rheumatoid arthritis: The signal importance of pain extent and fatigue, and comorbidity. *Arthritis Care & Research, 61*(5), 667-673.

Wolfe, F., & Michaud, K. (2011). The National Data Bank for Rheumatic Diseases: a multi-registry rheumatic disease data bank. *Rheumatology, 50*(1). doi:10.1093/rheumatology/keq155

WWAMI Rural Health Research Center. (2006). *Rural Urban Commuting Area Codes Data*. Retrieved November 27, 2018, from WWAMI Rural Health Research Center: http://depts.washington.edu/uwruca/ruca-approx.php

Acknowledgements

My achievements in this project were greatly aided by the following people:

- my capstone committee – Dr. Jane Meza, Dr. Lorena Baccaglini, and Dr. Kaleb Michaud;
- Rebecca Schumacher of FORWARD and Harlan Sayles of the UNMC College of Public Health and FORWARD for their assistance with data issues;
- Dr. Christopher Wichman and Dr. Gleb Haynatzki for lengthy conversations about analytical methods early in the process;
- my friends and colleagues Marcela Carvajal Suarez and Leah Swanson for giving me a glimpse of their own capstone processes so I had an idea of what to expect;
- my sister, Kristyn Clapp, for literally being my right hand when I realized at the last minute that my presentation needed reformatting beyond what I could handle with a broken thumb;
- my sister, Karolyn Grafel, for listening to all my woes along the way; and
- my parents, Kent Grafel and Mary Haller, for all the times they told me they were proud of me.

Appendix A – Codebook

Variable:	NDB Patient Key	**Type:**	Num
SAS Name:	PATKEY	**Format:**	16.0

Variable:	Survey Identifier	**Type:**	Num
SAS Name:	SURVEY	**Format:**	???

Value	Label
1	First
2	Last

Variable:	Number of Observations	**Type:**	Num
SAS Name:	NumObs	**Format:**	???

Notes: Number of observations in dataset for corresponding NDB Patient Key. Range is 1-2. A value of 1 indicates that the patient completed only a single survey. A value of 2 indicates that the patient completed more than 1 survey, and the dataset contains observations from the patient's first and last surveys.

Variable:	RA	**Type:**	Num
SAS Name:	RA	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Notes: Diagnosis of rheumatoid arthritis.

Variable:	SLE	**Type:**	Num
SAS Name:	SLE	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Notes: Diagnosis of systemic lupus erythematosus.

Variable:	Recruitment Method	**Type:**	Num (8)
SAS Name:	Recruit	**Format:**	RECRUITFMT

Value	Label	Value	Label
1	Arava	22	ARCK
2	30 Day NDB	23	UNMC/RAIN
3	Practices	24	Luggen
4	Wichita Databank	25	Bergman
6	Self Referral	26	MD Enroll 2008
7	FOCUS	27	BMS Website RALLY
8	Remicade-new start	28	Individual Site Enrollment
9	Remicade-old users	29	Cimzia Registry
10	Centocor Report Project	30	Edgerton
13	30day NoHAQ	31	AIRS
14	30day HAQ 2003	32	Walter Reed
16	International Website/Community	33	Soforo
17	Lupus-Community Project	34	FDR
18	HERO Followup study	35	UCSF RA Panel
19	Lupus-Harley/Oklahoma	37	UCSF Lupus
20	Katz Diagnosis Evaluation	38	International Dupuytren Data Bank
21	RALLY	39	UAB VERVE Study

Variable:	Simplified Recruitment Method	**Type:**	Num
SAS Name:	RECRUITSIMP	**Format:**	RECRUITSIMPFMT

Value	Label	Value	Label
1	Provider Referral	3	Drug Registries
2	Self-Enrolled	4	Other

Notes: Recategorization of RECRUIT into broader categories.
RECRUITSIMP = 1 when RECRUIT is 2, 3, 4, 17, 19, 22, 23, 24, 25, 26, 28, 30, 32, or 33.
RECRUITSIMP = 2 when RECRUIT is 6, 16, or 31.
RECRUITSIMP = 3 when RECRUIT is 1, 8, 9, 10, 27, or 29.
RECRUITSIMP = 4 when RECRUIT is any other value in the dataset.

Variable:	Date of Death	**Type:**	Num
SAS Name:	DEATHDAT	**Format:**	MMDDYY10

Notes: SAS date value for date of death.

Variable:	Questionnaire Type	**Type:**	Num (8)
SAS Name:	QTYPE	**Format:**	QTYPEFMT

Value	Label	Value	Label
10	CSQ-Comprehensive paper	1010	RAFib Paper
11	CSQ-Comprehensive telequest	1011	RAFib TQ
12	CSQ-Comprehensive Web	1012	RAFib Web
20	SSQ-Short paper	2010	OA Paper
21	SSQ-Short telequest	2011	OA TQ
22	SSQ-Short Web	2012	OA Web
30	BSQ-Brief paper	3010	Lupus Paper
31	BSQ-Brief telequest	3011	Lupus TQ
32	BSQ-Brief Web	3012	Lupus Web
512	Int l English-Comprehensive Web	4010	Gout Paper
730	Remicade-Brief paper	4012	Gout Web
731	Remicade-Brief telequest	6010	SpA Paper
810	Spanish-Comprehensive paper	6011	SpA Telequest
812	Spanish-Comprehensive Web		

Variable:	Education Level in Years	**Type:**	Num (8)
SAS Name:	EDLEVEL	**Format:**	BEST

Variable:	Education Category	**Type:**	Num (8)
SAS Name:	EDCAT	**Format:**	EDCATFMT

Value	Label	Value	Label
1	12 Years or Less	3	17 Years or More
2	13 – 16 Years		

Variable:	Race	**Type:**	Num (8)
SAS Name:	ETHORIG	**Format:**	ETHFMT

Value	Label	Value	Label
1	White, not of Hispanic origin	5	Hispanic
2	Black, not of Hispanic origin	6	Puerto Rican
3	Asian or Pacific Islander	7	Other
4	American Indian or Alaskan Native		

Variable:	Simplified Race	**Type:**	Num (8)
SAS Name:	ETHSIMP	**Format:**	ETHSIMPFMT

Value	Label	Value	Label
1	White, not of Hispanic origin	3	Other
2	Black, not of Hispanic origin		

Variable:	Sex of Patient	**Type:**	Num (8)
SAS Name:	SEX	**Format:**	SEXFMT

Value	Label
0	Female
1	Male

Variable:	Age of Patient	**Type:**	Num (8)
SAS Name:	AGE1 & AGE2	**Format:**	BEST

Variable:	Age Category	**Type:**	Num (8)
SAS Name:	AGECAT1 & AGECAT2	**Format:**	AGE4FMT

Value	Label	Value	Label
1	< 50	3	60 – 69
2	50 – 59	4	≥ 70

Variable:	Marital Status by Code	**Type:**	Num (8)
SAS Name:	MARITAL1 & MARITAL2	**Format:**	MARITALFMT

Value	Label	Value	Label
1	Never married	5	Widowed
2	Married	6	Widowed/remarried
3	Separated	7	Divorced/remarried
4	Divorced	8	Living together

Variable:	Simplified Marital Status	**Type:**	Num (8)
SAS Name:	MARSIMP1 & MARSIMP2	**Format:**	MARSIMPFMT

Value	Label
1	Single
2	Partnered

Notes: Dichotomous recategorization of marital status.
MARSIMP = 1 when MARITAL = 1, 3, 4, or 5.
MARSIMP = 2 when MARITAL = 2, 6, 7, OR 8.

Variable:	Total Annual Income	**Type:**	Num (8)
SAS Name:	INCOME1 & INCOME2	**Format:**	BEST

Value	Label	Value	Label
1	\$0 - \$9,999	7	\$60,000 - \$69,999
2	\$10,000 - \$19,999	8	\$70,000 - \$79,999
3	\$20,000 - \$29,999	9	\$80,000 - \$89,999
4	\$30,000 - \$39,999	10	\$90,000 - \$99,999
5	\$40,000 - \$49,999	11	\$100,000 - \$149,999
6	\$50,000 - \$59,999	12	\$150,000 or More

Variable:	Categorized Annual Income	**Type:**	Num (8)
SAS Name:	INCOMECAT1 & INCOMECAT2	**Format:**	INCOMECATFMT

Value	Label	Value	Label
1	Less Than \$30,000	3	\$60,000 or More
2	\$30,000 - \$59,999		

Notes: Recategorization of INCOME1 & INCOME2 variables.

Variable:	Body Mass Index	**Type:**	Num (8)
SAS Name:	BMI	**Format:**	BEST

Notes: Measured in kg/m^2.

Variable:	HAQ Disability Score	**Type:**	Num (8)
SAS Name:	HAQ1 & HAQ2	**Format:**	BEST

Notes: Range 0-3. Higher score indicates greater disability.

Variable:	HAQ II Score	**Type:**	Num (8)
SAS Name:	HAQII1 & HAQII2	**Format:**	BEST

Notes: Range 0-3. Higher score indicates greater disability.

Variable:	SF36 Physical Component Scale	**Type:**	Num (8)
SAS Name:	PCS1 & PCS2	**Format:**	BEST
Notes:	Range 0-100. Lower score indicates lower level of health.		

Variable:	SF36 Mental Component Scale	**Type:**	Num (8)
SAS Name:	MCS1 & MCS2	**Format:**	BEST
Notes:	Range 0-100. Lower score indicates lower level of health.		

Variable:	Rheumatic Disease Comorbidity Index (RDCI)	**Type:**	Num (8)
SAS Name:	COMOR1 & COMOR2	**Format:**	BEST

Variable:	Self-Assessed Health Status	**Type:**	Num (8)
SAS Name:	HEALTH1 & HEALTH2	**Format:**	BEST
Value	Label	Value	Label
1	Excellent	3	Fair
2	Good	4	Poor

Variable:	Simplified Health Status	**Type:**	Num (8)
SAS Name:	HEALTHSIMP1 & HEALTHSIMP2	**Format:**	HEALTHSIMPFMT
Value	Label	Value	Label
1	Excellent/Good		
2	Fair/Poor		
Notes:	Dichotomous recategorization of Self-Assessed Health Status. HEALTHSIMP is 1 when HEALTH is 1 or 2. HEALTHSIMP is 2 when HEALTH is 3 or 4.		

Variable:	Phase Number	**Type:**	Num (8)
SAS Name:	PHASE1 & PHASE2	**Format:**	PHASEFMT

Notes: Custom format is a text version of numeric value where 35 = ‘Phase 35’.
Mapping of dates to phases is found in Appendix B.

Variable:	Number of Drugs	**Type:**	Num (8)
SAS Name:	DRUGS1 & DRUGS2	**Format:**	BEST

Variable:	Number of DMARDs	**Type:**	Num (8)
SAS Name:	DMARDS1 & DMARDS2	**Format:**	BEST

Notes: DMARDs are disease-modifying anti-rheumatic drugs.

Variable:	Number of Biologics	**Type:**	Num (8)
SAS Name:	BIOLCNT1 & BIOLCNT2	**Format:**	BEST

Variable:	Analgesic Use	**Type:**	Num (8)
SAS Name:	ANALG1 & ANALG2	**Format:**	YNFMT

Notes: Includes acetaminophen products as well as opioids.

Variable:	Have CVA Problem Now	**Type:**	Num (8)
SAS Name:	STROKE1 & STROKE2	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Notes: Occurrence of stroke in past 6 months.

Variable:	Have CVO Problem Now	**Type:**	Num (8)
SAS Name:	HEART1 & HEART2	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Notes: Existence of heart condition other than heart attack in last 6 months.

Variable:	Employment Status	**Type:**	Num (8)
SAS Name:	EMPLOY1 & EMPLOY2	**Format:**	EMPFMT

Value	Label	Value	Label
0	Unemployed	4	Student
1	Paid work	5	Disabled
2	Retired	6	Working Part time
3	Housework	7	Other

Variable:	Have MI Problem Now	**Type:**	Num (8)
SAS Name:	MI1 & MI2	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Notes: Occurrence of heart attack in last 6 months (survey page 3).

Variable:	Cancer Problem Now	**Type:**	Num (8)
SAS Name:	CANCER1 & CANCER2	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Notes: Presence of cancer during last 6 months (survey page 3).

Variable:	Number of People Living in Household	**Type:**	Num (8)
SAS Name:	HOUSEHOLD1 & HOUSEHOLD2	**Format:**	BEST
Notes:			

Variable:	Flu Immunization in Current Phase	**Type:**	Num (8)
SAS Name:	FLU1 & FLU2	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Variable:	Ever Had Side Effect to Arthritis Medication	**Type:**	Num (8)
SAS Name:	SE1 & SE2	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Variable:	Infections in Current Phase	**Type:**	Num (8)
SAS Name:	INFXN1 & INFXN2	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Variable:	Zip Code (character)	**Type:**	Char (12)
SAS Name:	ZIP	**Format:**	N/A

Variable:	Zip Code (numeric)	**Type:**	Numeric
SAS Name:	ZIPN	**Format:**	5.0

Variable:	Diagnosis Group	**Type:**	Num (8)
SAS Name:	DX	**Format:**	DXFMT
Value	Label	Value	Label
1	RA	3	Other rheumatic diseases
2	SLE		

Notes: DX = 1 when RA = 1. DX = 2 when SLE = 1. DX -3 when RA = 0 and SLE = 0.

Variable:	Questionnaire Format	**Type:**	Num (8)
SAS Name:	QFORMAT1 & QFORMAT2	**Format:**	QFORMATFMT
Value	Label	Value	Label
1	Web	3	Telephone
2	Paper		

Notes: QFORMAT = 1 when QTYPE = 12, 22, 32, 512, 812, 1012, 2012, 3012, or 412.
QFORMAT = 2 when QTYPE = 10, 20, 30, 730, 810, 1010, 2010, 3010, 4010, or 6010.
QFORMAT = 3 when QTYPE = 11, 21, 31, 731, 1011, 2011, 3011 or 6011.

Variable:	Questionnaire Length	**Type:**	Num (8)
SAS Name:	QLEN1 & QLEN2	**Format:**	QLENFMT
Value	Label	Value	Label
1	Comprehensive	3	Brief
2	Short		

Notes: QLEN = 1 when QTYPE = 10, 11, 12, 512, 810, 812, 1010, 1011, 1012, 2010, 2011, 2012, 3010, 3011, 3012, 4010, 4012, 6010, or 6011.
QLEN = 2 when QTYPE = 20, 21, or 22.
QLEN = 3 when QTYPE = 30, 31, 32, 730, or 731.

Variable:	Death Phase	**Type:**	Num (8)
SAS Name:	DEATHPHASE	**Format:**	PHASEFMT

Notes: Phase during which DEATHDAT occurs. Mapped according to schedule of phases in Appendix B.

Variable:	Censoring Status	**Type:**	Num (8)
SAS Name:	CENSOR	**Format:**	CENSORFMT

Value	Label
0	Censored
1	Not Censored

Notes: CENSOR = 0 when DEATHPH = PHASE2 + 1.
CENSOR = 0 when PHASE2 = 39.
CENSOR = 1 if neither condition is met.

Variable:	Duration of Participation	**Type:**	Num (8)
SAS Name:	DURATION	**Format:**	BEST

Notes: Calculated as: $DURATION = PHASE2 - PHASE1$

Variable:	Dropout at Less than 2 Years	**Type:**	Num (8)
SAS Name:	DROPOUT	**Format:**	YNFMT

Value	Label
0	No
1	Yes

Notes: DROPOUT = 0 when DURATION ≥ 4.
DROPOUT = 1 when DURATION < 4.

Variable:	Change in HAQ Disability Score	**Type:**	Num (8)
SAS Name:	HAQCHG	**Format:**	BEST

Notes: Calculated as: $HAQCHG = HAQ2 - HAQ1$
Range: -3 to +3.

Variable:	Change in HAQ II Score	**Type:**	Num (8)
SAS Name:	HAQIICHG	**Format:**	BEST
Notes:	Calculated as: $HAQIICHG = HAQII2 - HAQII1$ Range: -3 to +3.		

Variable:	Change in SF36 Physical Component Scale	**Type:**	Num (8)
SAS Name:	PCSCHG	**Format:**	BEST
Notes:	Calculated as: $PCSCHG = PCS2 - PCS1$ Range: -100 to +100.		

Variable:	Change in SF36 Mental Component Scale	**Type:**	Num (8)
SAS Name:	MCSCHG	**Format:**	BEST
Notes:	Calculated as: $MCSCHG = MCS2 - MCS1$ Range: -100 to +100.		

Variable:	Change in RDCI	**Type:**	Num (8)
SAS Name:	COMORCH	**Format:**	BEST
Notes:	Calculated as: $COMORCHG = COMOR2 - COMOR1$ Range: -9 to +9.		

Variable:	Change in Number of Drugs	**Type:**	Num (8)
SAS Name:	DRUGCHG	**Format:**	BEST
Notes:	Calculated as: $DRUGCHG = DRUGS2 - DRUGS1$		

Variable:	Change in Number of DMARDs	**Type:**	Num (8)
SAS Name:	DMARDCHG	**Format:**	BEST
Notes:	Calculated as: $DMARDCHG = DMARDS2 - DMARDS1$		

Variable:	Change in Number of Biologic Drugs	**Type:**	Num (8)
SAS Name:	BIOLCHG	**Format:**	BEST

Notes: Calculated as: $BIOLCHG = BIOLCNT2 - BIOLCNT1$

Variable:	Change in Self-Assessed Health Status	**Type:**	Num (8)
SAS Name:	HEALTHCHG	**Format:**	BEST

Notes: Calculated as: $HEALTHCHG = HEALTH2 - HEALTH1$
Range: -3 to +3.

Variable:	RUCA Code	**Type:**	Num (8)
SAS Name:	RUCACOD	**Format:**	BEST

Notes: Index describing degree of urbanization. RUCACOD is mapped to ZIP using tables from the Rural Health Research Center. Variable takes discrete, nominal values ranging from 1.0 to 10.6.

Variable:	RUCA Category	**Type:**	Num (8)
SAS Name:	RUCACAT	**Format:**	RUCACATFMT

Value	Label		
1	Urban	3	Small rural
2	Large rural	4	Isolated

Notes: Groupings of RUCA codes defined by the Rural Health Research Center.
RUCACAT = 1 when RUCACOD = 1.0, 1.1, 2.0, 2.1, 3.0, 4.1, 5.1, 7.1, 8.1, or 10.1.
RUCACAT = 2 when RUCACOD = 4.0, 4.2, 5.0, 5.2, 6.0, or 6.1.
RUCACAT = 3 when RUCACOD = 7.0, 7.2, 7.3, 7.4, 8.0, 8.2, 8.3, 8.4, 9.0, 9.1, or 9.2.
RUCACAT = 4 when RUCACOD = 10.0, 10.2, 10.3, 10.4, 10.5, or 10.6.

Variable:	Simplified RUCA Category	**Type:**	Num (8)
SAS Name:	RUCASIMP	**Format:**	RUCASIMPFMT

Value	Label
1	Urban
2	Rural or Isolated

Notes: Simplified dichotomization of RUCA Category.
RUCASIMP is 1 when RUCACAT is 1.
RUCASIMP is 2 when RUCACAT is 2, 3, or 4.

Appendix B – FORWARD Phases

An index of the phases identified in FORWARD data is given in Table B1. Survey periods begin in January and July of each year. Surveys are typically completed within the 6-month period beginning with January or July, though a small number of surveys in the dataset were returned slightly later. The review period is the 6-month period prior to the survey's release. Survey questions predominantly ask about patient experience during this time. The phase number refers to the review period, not the period during which the survey is released.

Table B1

FORWARD Phases

Phase	Review Period	Survey Date	Phase	Review Period	Survey Date
35	January - June 1998	July 1998	55	January - June 2008	July 2008
36	July - December 1998	January 1999	56	July - December 2008	January 2009
37	January - June 1999	July 1999	57	January - June 2009	July 2009
38	July - December 1999	January 2000	58	July - December 2009	January 2010
39	January - June 2000	July 2000	59	January - June 2010	July 2010
40	July - December 2000	January 2001	60	July - December 2010	January 2011
41	January - June 2001	July 2001	61	January - June 2011	July 2011
42	July - December 2001	January 2002	62	July - December 2011	January 2012
43	January - June 2002	July 2002	63	January - June 2012	July 2012
44	July - December 2002	January 2003	64	July - December 2012	January 2013
45	January - June 2003	July 2003	65	January - June 2013	July 2013
46	July - December 2003	January 2004	66	July - December 2013	January 2014
47	January - June 2004	July 2004	67	January - June 2014	July 2014
48	July - December 2004	January 2005	68	July - December 2014	January 2015
49	January - June 2005	July 2005	69	January - June 2015	July 2015
50	July - December 2005	January 2006	70	July - December 2015	January 2016
51	January - June 2006	July 2006	71	January - June 2016	July 2016
52	July - December 2006	January 2007	72	July - December 2016	January 2017
53	January - June 2007	July 2007	73	January - June 2017	July 2017
54	July - December 2007	January 2008	74	July - December 2017	January 2018

Appendix C – Power Analysis

Power analysis was completed for logistic regression models (Aim 1) and survival models (Aim 2). Both analyses were based on 90% power (β = 0.1), 0.05 significance (α), and a sample size of 5,000 (*n*). A total sample size of approximately 50,000 was expected; however, the 3 diagnosis groups were anticipated to vary widely in size, with *n* = 5,000 anticipated to be the smallest sample size. Therefore, these power analyses were considered conservative since the sample sizes for some groups were larger. Preliminary analysis of the data, however, revealed that the SLE group contained only 2,752 subjects. Power for the SLE models is addressed below.

Logistic Regression Models

For a binary predictor variable in a logistic regression model, Tables C1 and C2 display the lowest detectable odds ratios given a set of varying parameters: predictor variable distribution (percentage of sample with X=1), response probability (P_0), and correlation between predictor variables (R^2). Power was constant at 90% and significance at 0.05 for all scenarios.

Table C1

Odds Ratios Detectable at 90% Power and α = 0.05

20% of sample with X = 1			
	R^2		
P_0	**0.3**	**0.5**	**0.7**
0.15	1.431	1.522	1.702
0.30	1.337	1.407	1.549
0.45	1.315	1.383	1.520

40% of sample with X = 1			
	R^2		
P_0	**0.3**	**0.5**	**0.7**
0.15	1.046	1.060	1.030
0.30	1.036	1.029	1.038
0.45	1.023	1.027	1.035

For a binary predictor variable, assume that 40% of the sample has the characteristic (X=1) and the probability of dropout for that group is 0.3. If the highest pairwise correlation between the predictor and the other variables in the model is 0.5, then the lowest detectable odds ratio is 1.036 (Hsieh, Block, & Larsen 1998).

Survival Models

Power analysis for the survival model assumed an 80% event rate. This was based on the information that approximately 50,000 patients have participated in FORWARD surveys throughout its lifespan, and approximately 10,000 patients participate currently. Table C2 below displays the lowest detectable hazard ratios given varying predictor variable standard distribution (SD) and correlation (R^2). Power was constant at 90% and significance at 0.05 for all scenarios.

Table C2

Hazard Ratios Detectable at 90% Power and α = 0.05

	R^2		
P_0	**0.3**	**0.5**	**0.7**
0.15	1.431	1.522	1.702
0.30	1.337	1.407	1.549
0.45	1.315	1.383	1.520

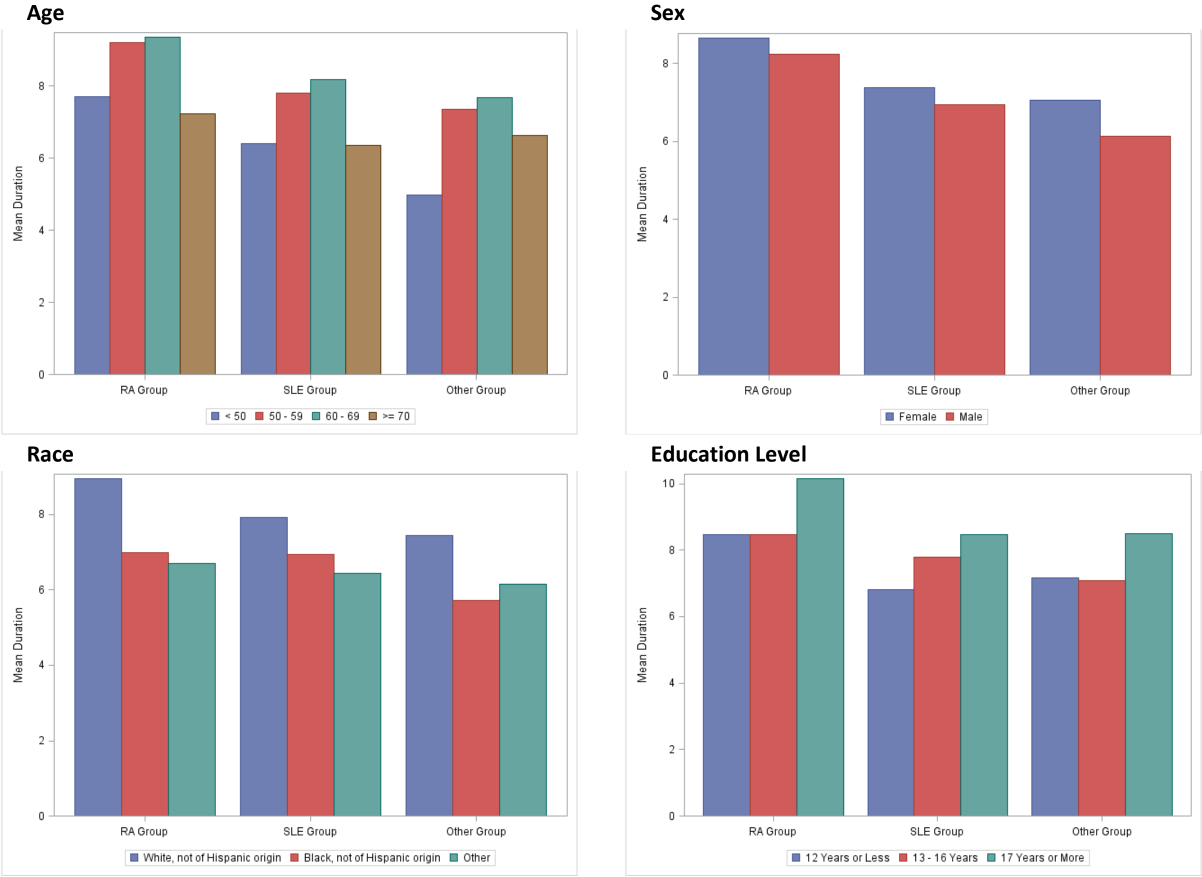
For a continuous predictor variable assume a standard deviation of 10. If the highest pairwise correlation between the predictor and the other variables in the model is 0.3, then the lowest detectable hazard ratio is 1.006 (Hsieh & Lavori 2000; Schoenfeld 1983).

Appendix D – Duration by Selected Patient Characteristics

Duration patterns were assessed separately for each patient characteristic listed in Table 2. Visual representations of the findings are given in Figure D1. This initial assessment was informative only, as each of these predictors was considered for the models regardless of apparent effect on duration.

Figure D1

Mean Duration by Patient Characteristics for All Groups



Age

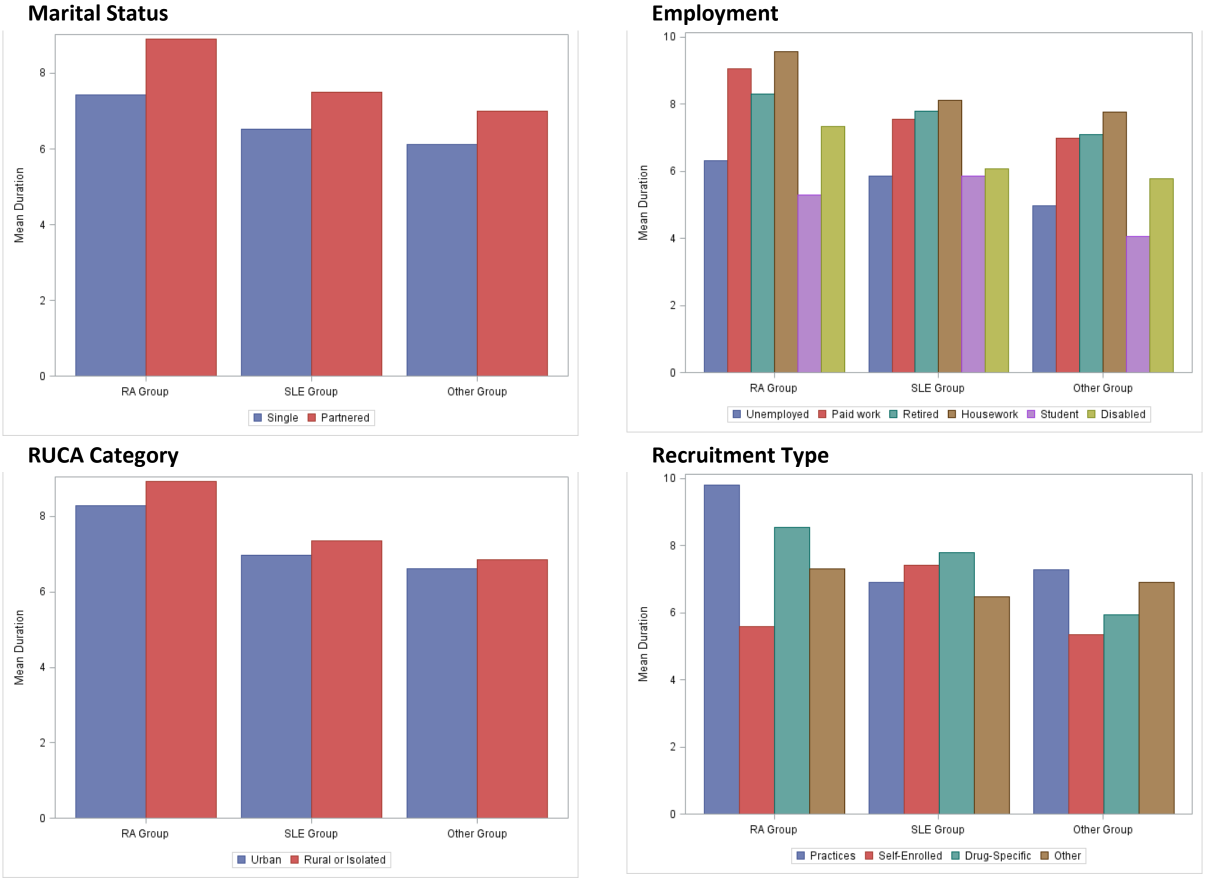
Sex

Race

Education Level

Figure D1

Mean Duration by Patient Characteristics for All Groups

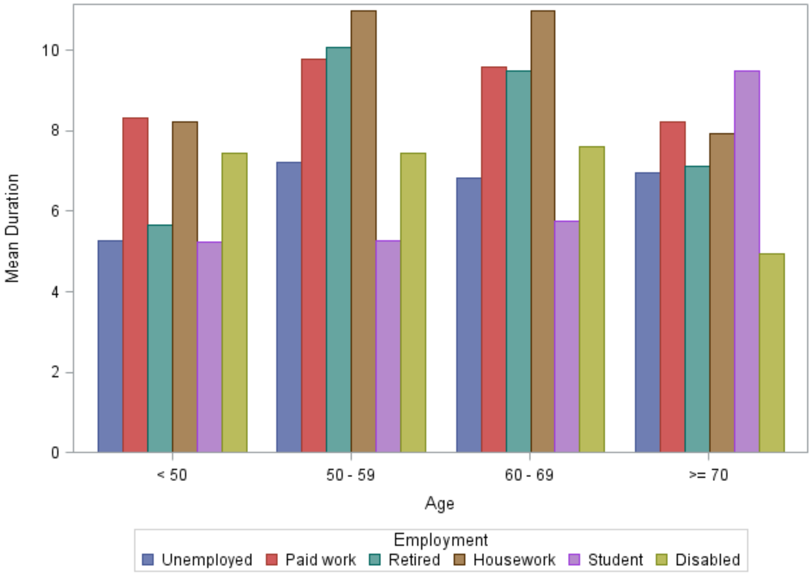


Note. All characteristics are assessed as of the first survey.

Further analysis was conducted jointly for pairings of these characteristics. The pairwise assessments were more instructive than the single characteristic, as these plots held the potential to reveal possible interactions in the multivariable logistic regression models. Differences in the patterns of vertical bars between clusters might indicate that the effect of one characteristic on duration was dependent on the level of the paired characteristic. Each possible pairing of characteristics was assessed, separately for each diagnosis group. Plots that displayed notable variation are shown in Figures D2, D3, and D4. Such variation was most commonly seen when evaluating characteristics with many levels, particularly the employment variable. Variation was also more common in plots for the SLE and other rheumatic disease groups than for the much larger RA group. These observations suggest that any apparent variation is likely an effect of small cell counts when the two characteristics are cross-tabulated, which result in less regression toward a mean outcome. As such, perceived differences may not be strong evidence of an interaction. After employing additional methods to select interaction terms, the only pairing shown below that was used in a full model was that of age and employment in the other rheumatic diseases group. This interaction was eliminated during selection of the final model.

Figure D2

Mean Duration by Age and Employment for RA Group



Note. Both characteristics are assessed as of the first survey.

Figure D3

Mean Duration by Selected Characteristic Pairs for SLE Group

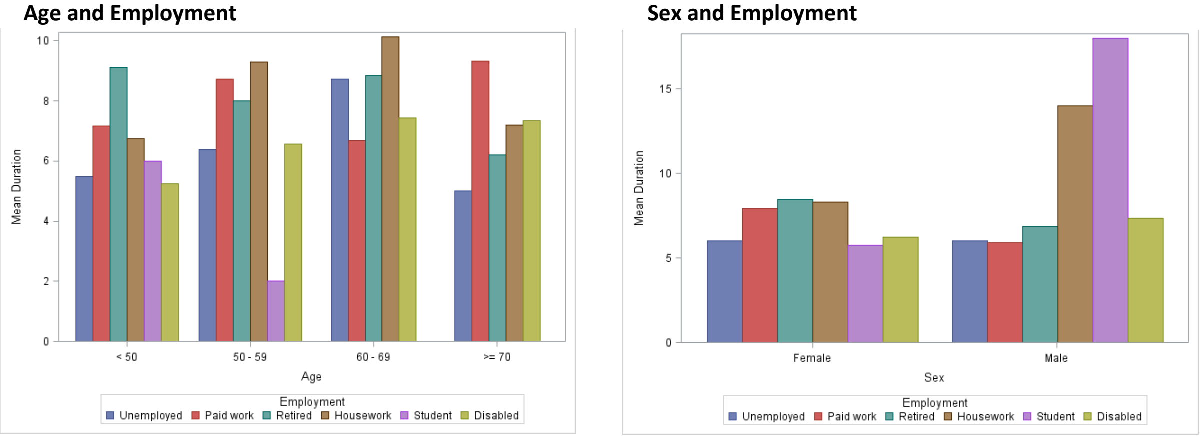
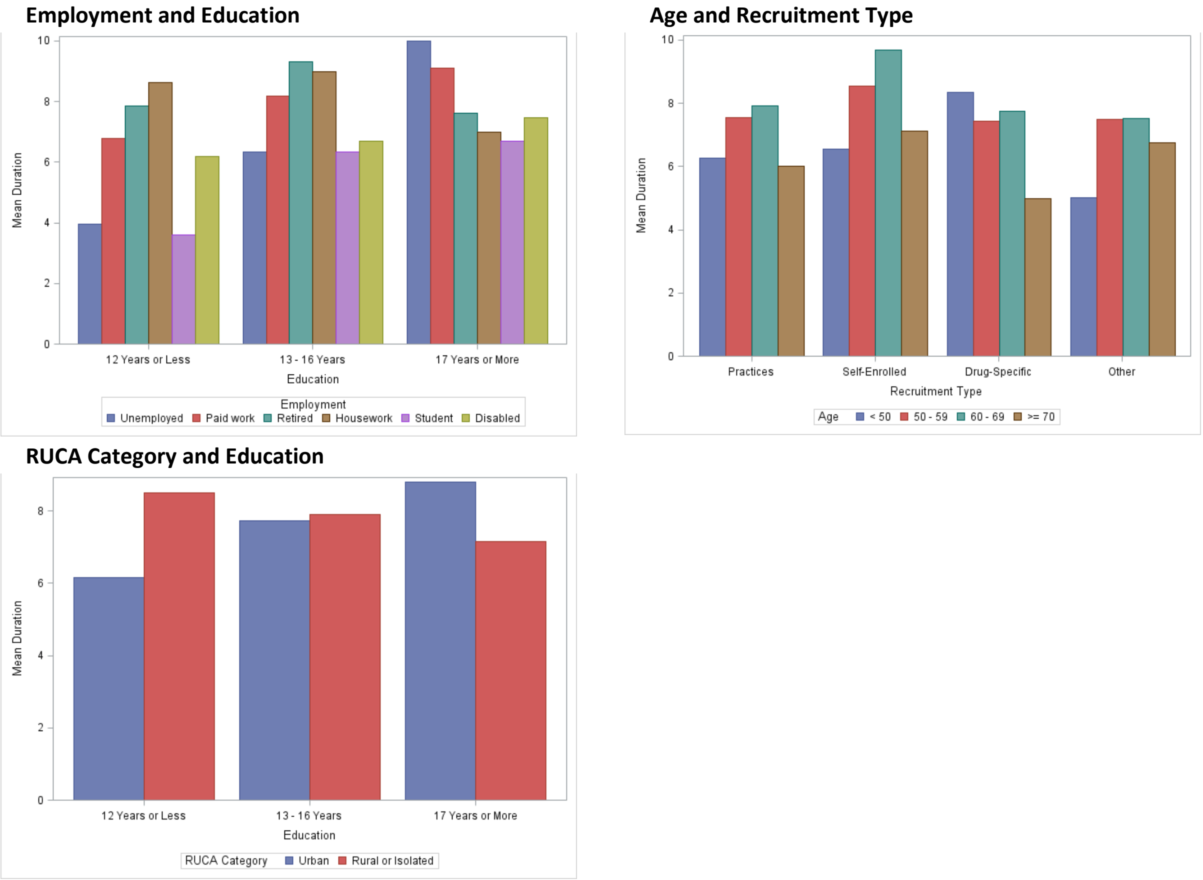


Figure D3

Mean Duration by Selected Characteristic Pairs for SLE Group



Employment and Education

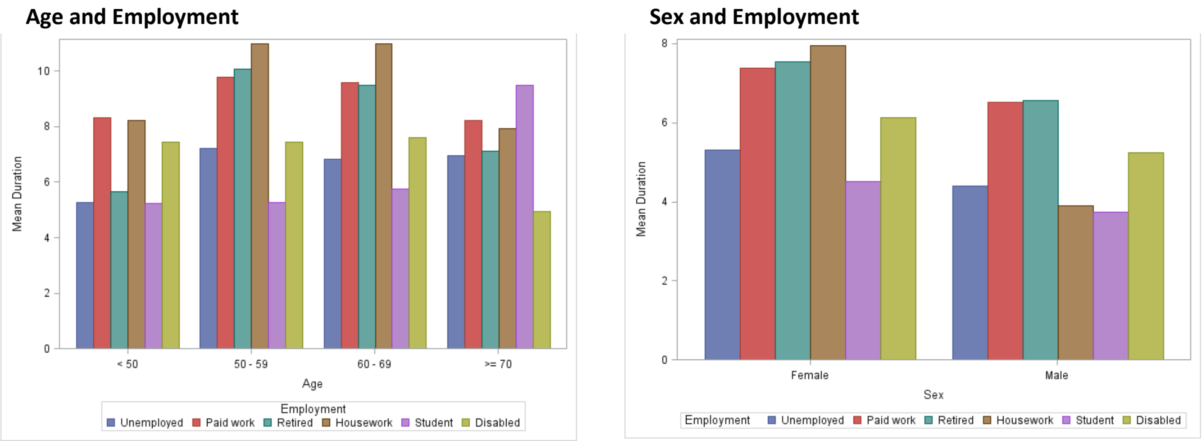
Age and Recruitment Type

RUCA Category and Education

Note. All characteristics are assessed as of the first survey.

Figure D4

Mean Duration by Selected Characteristic Pairs for Other Rheumatic Diseases Group

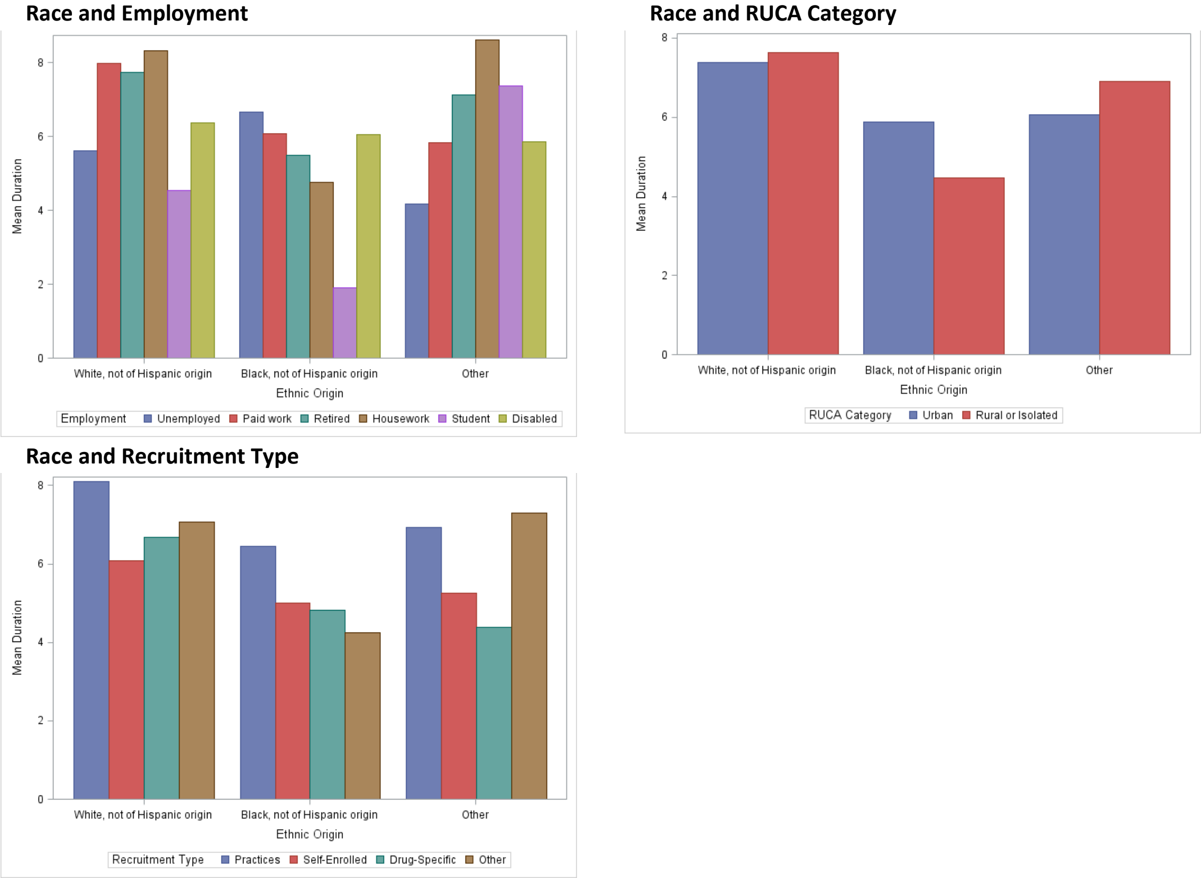


Age and Employment

Sex and Employment

Figure D4

Mean Duration by Selected Characteristic Pairs for Other Rheumatic Diseases Group



Note. All characteristics are assessed as of the first survey.

Appendix E – Logistic Regression Model for RA Group

Results of simple and multiple logistic regressions for patients with RA are presented in the following tables. Table E1 contains specifications for univariable models that were constructed separately for each predictor having adequate data and heterogeneity. In cases of categorical variables with more than 2 levels, a set of indicator terms was used in place of the original variable.

Table E1

Univariable Logistic Regression Results for RA Group

Parameter	*b*	SE	*p*	OR [95% CI]	Notes
Sex (Male)	0.0674	0.0265	0.011	1.07 [1.02 – 1.13]	Reference group: female. Dropped from consideration due to odds ratio close to 1.
Education level (Year)	-0.0462	0.0046	< 0.001	0.96 [0.95 – 0.96]	
Race (Black)	0.4213	0.0493	< 0.001	1.52 [1.38 – 1.68]	Reference group: White.
Race (Other)	0.5070	0.0441	< 0.001	1.66 [1.52 – 1.81]	
Recruitment (Provider Referral)	-0.2917	0.0330	< 0.001	0.75 [0.70 – 0.80]	Reference group: other recruitment methods.
Recruitment (Self-Enrolled)	0.0540	0.0402	0.179	1.06 [0.98 – 1.14]	
Recruitment (Drug Registries)	-0.2248	0.0331	< 0.001	0.80 [0.75 – 0.85]	
Age (Years)	- 0.0072	0.0008	< 0.001	0.99 [0.99 – 0.99]	OR for unit of 10 years was 0.93 (95% CI 0.92, 0.95). Dropped from consideration due to odds ratio close to 1.
Marital Status (Single)	- 0.2265	0.0238	< 0.001	0.80 [0.76 – 0.84]	Reference group: partnered.
Employment (Housework)	- 0.3300	0.0426	< 0.001	0.72 [0.66 – 0.78]	Reference group: disabled.
Employment (Paid)	- 0.1920	0.0337	< 0.001	0.83 [0.77 – 0.88]	
Employment (Retired)	-0.2572	0.0351	< 0.001	0.78 [0.72 – 0.83]	
Employment (Student)	0.4019	0.1221	0.001	1.50 [1.18 – 1.90]	
Employment (Unemployed)	0.2296	0.0677	0.001	1.26 [1.10 – 1.44]	
Income ($30,000 - $59,999)	- 0.1809	0.0293	< 0.001	0.83 [0.79 – 0.88]	Reference group: income ≤ $30,000. Dropped from consideration due to lack of evidence for categorization being meaningful.
Income ($60,000 or More)	- 0.1932	0.0291	< 0.001	0.32 [0.78 – 0.88]	
RUCA Category (Urban)	- 0.1250	0.0250	< 0.001	0.88 [0.84 – 0.93]	Reference group: rural/isolated.
HAQ	0.1721	0.0161	< 0.001	1.19 [1.15 – 1.23)	
PCS	- 0.0115	0.0011	< 0.001	0.99 [0.99 – 0.99]	Odds ratio for a change of 10 units was 0.89 [0.87, 0.91]. Dropped from consideration due to significant correlation with HAQ.
RDCI	0.0899	0.0071	< 0.001	1.09 [1.08 – 1.11]	
Health Status (Excellent/Good)	0.4168	0.0228	< 0.001	1.52 [1.45 – 1.59]	Reference group: fair/poor.
Number of Drugs	- 0.0201	0.0026	< 0.001	0.98 [0.98, 0.99]	Dropped from consideration due to odds ratio close to 1.
Number of DMARDs	- 0.1392	0.0116	< 0.001	0.87 [0.85 – 0.89]	Dropped from consideration due to highly non-normal distribution.
Analgesic Use (Yes)	- 0.0051	0.0221	0.816	1.00 [0.95 – 1.04]	
Heart Problem (Yes)	0.1843	0.0420	< 0.001	1.20 [1.11 – 1.31]	

Note. All variables are assessed as of first survey. RA = rheumatoid arthritis; RUCA = Rural Urban Commuting Area Codes; HAQ = health assessment questionnaire; PCS = physical component scale; RDCI = rheumatic disease comorbidity index; DMARDs = disease-modifying anti-rheumatic drugs.

Results of these univariable models were used to select predictors for the full multivariable model. For continuous and dichotomous variables, criterium for inclusion was $p < 0.05$ in the relevant univariable model. For multi-level categorical variables, the criterium was $p < 0.05$ for all indicator variables or, in cases where a subset of indicator terms was significant, $p < 0.05$ for the Type 3 analysis of effect. Qualifying continuous predictors were assessed for multicollinearity. HAQ and PCS scores were found to be moderately correlated (r = - 0.744, $p < 0.001$). PCS was dropped in favor of HAQ due to more missing values for PCS (20.8% missing PCS vs. 13.9% missing HAQ). Due to an excess of qualifying predictors, age, income, and number of DMARDs were excluded from the full model based on subjective assessment. Selection proceeded by sequential removal of predictors to obtain the final model.

Results of the full and final multivariable models for the RA group are presented in Table E2, and the series of steps taken to select the final model is given in Table E3. Removal of only one non-significant predictor was required to obtain the final model for the RA group.

Table E2

Multivariable Logistic Regression Results for RA Group

	Full Model			Final Model			
Parameter	***b***	**SE**	***p***	***b***	**SE**	***p***	**OR [95% CI]**
Intercept	0.3812	0.1405	0.007	0.3816	0.1405	0.007	N/A
Race (Black)	0.3349	0.0625	< 0.001	0.3350	0.0625	< 0.001	1.40 [1.24 - 1.58]
Race (Other)	0.3827	0.0547	< 0.001	0.3828	0.0547	< 0.001	1.47 [1.32 - 1.63]
Marital Status (Single)	-0.1405	0.0291	< 0.001	-0.1404	0.0291	< 0.001 [a]	0.87 [0.82 - 0.92]
Education Level (Years)	-0.0529	0.0059	< 0.001	-0.0529	0.0059	< 0.001	0.95 [0.94 - 0.96]
Recruitment (Provider Referral)	-0.2630	0.0412	< 0.001	-0.2632	0.0412	< 0.001 [a]	0.77 [0.71 - 0.83]
Recruitment (Self-Enrolled)	0.0226	0.0484	0.640	0.0226	0.0484	0.640 [a]	1.02 [0.93 - 1.12] [b]
Recruitment (Drug Registries)	-0.2215	0.0419	< 0.001	-0.2217	0.0419	< 0.001 [a]	0.80 [0.74 - 0.87]
Employment (Housework)	-0.0321	0.0529	0.544	-0.0321	0.0529	0.544 [a]	0.97 [0.87 - 1.07] [b]
Employment (Paid Work)	0.1393	0.0451	0.002	0.1394	0.0451	0.002 [a]	1.15 [1.05 - 1.26]
Employment (Retired)	0.0164	0.0445	0.713	0.0162	0.0445	0.716 [a]	1.02 [0.93 - 1.11] [b]
Employment (Student)	0.6088	0.1398	< 0.001	0.6089	0.1398	< 0.001 [a]	1.84 [1.40 - 2.42]
Employment (Unemployed)	0.3143	0.0794	< 0.001	0.3145	0.0794	< 0.001 [a]	1.37 [1.17 - 1.60]
RUCA Category (Urban)	-0.1086	0.0315	0.001	-0.1087	0.0315	0.001	0.90 [0.84 - 0.95]
HAQ Score	-0.2987	0.0702	< 0.001	-0.2984	0.0702	< 0.001	0.74 [0.65 - 0.85)
RDCI	0.0545	0.0099	< 0.001	0.0537	0.0093	< 0.001	1.06 [1.04 - 1.07]
Health Status (Excellent/Good)	0.0881	0.0563	0.118	0.0881	0.0563	0.117 [c]	1.09 [0.98 - 1.22] [b]
Heart Problem (Yes)	-0.0134	0.0549	0.807	--	--	--	--

Table E2

Multivariable Logistic Regression Results for RA Group

	Full Model			Final Model			
Parameter	***b***	**SE**	***p***	***b***	**SE**	***p***	**OR [95% CI]**
Infection (Yes)	0.1164	0.0293	< 0.001	0.1165	0.0293	< 0.001	1.12 [1.06 - 1.19]
Infection (Yes) * HAQ Score	0.2029	0.0440	< 0.001	0.2028	0.0440	< 0.001	1.22 [1.12 - 1.34]

Note. Reference categories for categorical variables are given in Table E1. All values are assessed as of the first survey. RA = rheumatoid arthritis; RUCA = Rural Urban Commuting Area Codes; HAQ = health assessment questionnaire; RDCI = rheumatic disease comorbidity index.

[a] Type 3 analysis of effect indicated a significant overall effect for these predictors (recruitment: $p < 0.001$; employment: $p < 0.001$). [b] Confidence interval is inclusive of 1. Odds ratio may be informative but should not be considered definitive. [c] The main effect of health status was retained in the model due to involvement in a significant interaction.

Table E3

Logistic Regression Model Selection for RA Group

#	Description	Action Taken	AIC
1	Full Model		32169.76
2	Final Model	Removed Heart Problem	32167.82
3	*Exploratory Model*	*Removed HAQ * Health Status*	*32187.11*

Note. A chi-square test comparing Models 2 and 3 indicated that the model containing the interaction term was a significantly better fit than the reduced model ($X^2 = 19.297$, $p < 0.001$). RA = rheumatoid arthritis; HAQ = health assessment questionnaire.

Indicator terms for multilevel categorical variables required selection of a single reference group; however, comparisons among other levels of the predictors are of interest. To this end, odds ratios comparing selected levels of the race, recruitment, and employment variables are presented in Table E4. Given greater interest in the groups that are at higher risk of failure, all comparisons are made in the direction resulting in an odds ratio greater than 1.

Table E4

Selected Pairwise Odds Ratios for the RA Group

Race		Employment	
Comparison	**OR [95% CI]**	**Comparison**	**OR [95% CI]**
Black vs. White	1.40 [1.24, 1.58]	Paid Work vs. Disabled	1.15 [1.05, 1.26]
Other vs. White	1.47 [1.32, 1.63]	Paid Work vs. Retired	1.13 [1.06, 1.21]
Other vs. Black	1.05 [0.90, 1.23] [a]	Unemployed vs. Paid Work	1.19 [1.03, 1.38]
Recruitment		Retired vs. Disabled	1.02 [0.93, 1.11] [a]
Comparison	**OR [95% CI]**	Unemployed vs. Retired	1.35 [1.16, 1.56]
Self-Enrolled vs. Provider Referral	1.33 [1.23, 1.44]		
Drug Registries vs. Provider Referral	1.04 [0.98, 1.11] [a]		
Self-Enrolled vs. Drug Registries	1.28 [1.18, 1.39]		

Note. RA = rheumatoid arthritis.

[a] Confidence interval is inclusive of 1. Odds ratio may be informative but should not be considered not definitive.

Appendix F – Logistic Regression Model for SLE Group

Results of univariable logistic regression models for the SLE group are presented in Table F1. Refer to Appendix E for more information.

Table F1

Univariable Logistic Regression Results for SLE Group

Parameter	*b*	SE	*p*	OR [95% CI]	Notes
RA	- 0.1648	0.0974	0.091	0.85 [0.70 – 1.03]	
Sex (Male)	0.1901	0.1634	0.245	1.21 [0.88 – 1.67]	
Education Level (Years)	-0.0520	0.0180	0.004	0.95 [0.92, 0.98]	
Race (Black)	0.1536	0.1171	0.190	1.17 [0.93 – 1.47]	
Race (Other)	0.2056	0.1337	0.124	1.23 [0.95 – 1.60]	
Recruitment (Provider Referral)	- 0.9757	0.3886	0.012	0.38 [0.18 – 0.81]	Type 3 analysis of effect indicated a significant overall effect for the categorical variable (*p* = 0.005).
Recruitment (Self-Enrolled)	0.0717	0.0810	0.376	1.07 [0.92 – 1.26]	
Recruitment (Drug Registries)	- 0.3143	0.1435	0.029	0.73 [0.55 – 0.97]	
Age	- 0.0108	0.0029	< 0.001	0.99 [0.99 – 0.99]	
Marital Status (Single)	- 0.1817	0.0789	0.021	0.83 [0.71 – 0.97]	
Employment (Housework)	- 0.2469	0.1371	0.072	0.78 [0.60 – 1.02]	Type 3 analysis of effect indicated a significant overall effect for the categorical variable (*p* < 0.001).
Employment (Paid)	- 0.3501	0.0982	< 0.001	0.71 [0.58 – 0.85]	
Employment (Retired)	- 0.4738	0.1308	< 0.001	0.62 [0.48 – 0.81]	
Employment (Student)	0.0564	0.2657	0.832	1.06 [0.63 – 1.78]	
Employment (Unemployed)	0.0724	0.1870	0.699	10.8 [0.75 – 1.56]	
Income ($30,000 - $59,999)	- 0.0781	0.1030	0.448	0.93 [0.76 – 1.13]	Type 3 analysis of effect indicated a significant overall effect for the categorical variable (*p* < 0.001).
Income ($60,000 or More)	- 0.4036	0.0958	< 0.001	0.67 [0.55 – 0.81]	
RUCA Category (Urban)	- 0.0780	0.0898	0.385	0.93 [0.78 – 1.10]	
HAQ	0.1459	0.0572	0.011	1.16 [1.03 – 1.30]	
PCS	- 0.0097	0.0037	0.008	0.99 [0.98 – 0.99]	Odds ratio for a change of 10 units was 0.91 [0.85, 0.98].
RDCI	0.0132	0.0205	0.520	1.01 [0.97 – 1.06]	
Health Status (Excellent/Good)	0.3162	0.0791	< 0.001	1.37 [1.18 – 1.60]	
Number of Drugs	- 0.0277	0.0082	< 0.001	0.97 [0.96 – 0.99]	
Number of DMARDs	- 0.1637	0.0497	0.0010	0.85 [0.77 – 0.94]	
Analgesic Use (Yes)	0.0137	0.0782	0.861	1.01 [0.87 – 1.18]	
Heart Problem (Yes)	0.0220	0.1168	0.851	1.02 [0.81 – 1.29]	
Infection (Yes)	0.0216	0.0790	0.785	1.02 [0.88 – 1.19]	

Note. All variables are assessed as of first survey. SLE = systemic lupus erythematosus; RUCA = Rural Urban Commuting Area Codes; HAQ = health assessment questionnaire; PCS = physical component scale; RDCI = rheumatic disease comorbidity index; DMARDs = disease-modifying anti-rheumatic drugs.

As with the RA group, HAQ and PCS scores were moderately correlated (*r* = - 0.752, *p* < 0.001). In the case of the SLE group, however, missing rates for the two variables were similar (HAQ 14.0%, PCS 13.3%). Given that the correlation coefficient was borderline to the threshold set as warranting action (*r*

= 0.8), several exploratory models were considered to assess whether parameter estimate for each variable was greatly affected by the presence or absence of the collinear predictor. These exploratory models contained HAQ score, PCS score, all other qualifying predictors and 2 interaction terms that had been previously identified as potentially relevant. Parameter estimates for these models are given in Table F2. The parameter estimate for HAQ changed in both magnitude and direction depending on the presence of PCS, whereas the parameter estimate for PCS remained essentially the same. Further, although the effect of neither predictor was significant, the *p*-values for PCS was much lower than those for HAQ. As a result, PCS score was included in the full multivariable model, and HAQ score was eliminated.

Table F2

Evaluation of HAQ and PCS Collinearity Effect

	HAQ		**PCS**	
Model	***b***	***p***	***b***	***p***
Model containing both HAQ and PCS	- 0.0445	0.685	- 0.0139	0.054
Model containing HAQ only	0.0888	0.383	--	--
Model containing PCS only	--	--	- 0.0089	

Note. Models also contained all covariates that appeared in the full model.
HAQ = health assessment questionnaire; PCS = physical component scale.

All remaining variables that met the selection criteria were included in the full model for the SLE group. Specifications of the full model and the final model are given in Table F3, and the series of steps taken to obtain the final model is given in Table F4. A significant interaction effect between self-assessed health status and employment category was removed from the model in the interest of parsimony. The model containing the interaction term was not a significantly better fit than the reduced model (X^2 = 8.143, p = 0.149).

Table F3

Multivariable Logistic Regression Results for SLE Group

	Full Model			Final Model			
Parameter	***b***	**SE**	***p***	***b***	**SE**	***P***	**OR [95% CI]**
Intercept	0.3569	0.7247	0.622	0.8655	0.3127	0.006	N/A
Age (Years)	- 0.0119	0.0048	0.013	-0.0122	0.0041	0.003	0.99 [0.98 - 1.00]
Marital Status (Single)	0.0599	0.1739	0.731	--	--	--	--
Education Level (Years)	- 0.0219	0.0218	0.315	--	--	--	--
Recruitment (Provider Referral)	- 0.2132	0.1519	0.160	-0.2026	0.1306	0.121 [a]	0.82 [0.63 - 1.05] [b]
Recruitment (Self-Enrolled)	- 0.3399	0.157	0.031	-0.4250	0.1400	0.002 [a]	0.65 [0.50 - 0.86]
Recruitment (Drug Registries)	- 0.3713	0.2015	0.065	-0.5069	0.1863	0.007 [a]	0.60 [0.42 - 0.87]
Employment (Housework)	1.9345	0.6502	0.003	-0.2387	0.1565	0.127 [a]	0.79 [0.58 - 1.07] [b]
Employment (Paid work)	0.9331	0.5380	0.083	-0.3542	0.1235	0.004 [a]	0.70 [0.55 - 0.89]
Employment (Retired)	0.0705	0.6773	0.917	-0.3654	0.1618	0.024 [a]	0.69 [0.51 - 0.95]
Employment (Student)	- 1.4369	1.208	0.234	-0.3369	0.3019	0.264 [a]	0.71 [0.40 - 1.29] [b]
Employment (Unemployed)	0.3926	0.874	0.653	0.0687	0.2001	0.731 [a]	1.07 [0.72 - 1.59] [b]
Income ($30,000 - $59,999)	0.2262	0.4124	0.583	-0.0014	0.1104	0.990 [a]	1.00 [0.80 - 1.24] [b]
Income ($60,000 or More)	0.3004	0.4774	0.529	-0.2682	0.1074	0.012 [a]	0.76 [0.62 - 0.94]
Health Status (Excellent/Good)	0.7090	0.2619	0.007	0.2251	0.0931	0.016	1.25 [1.04 - 1.50]
PCS Score	- 0.0094	0.0059	0.114	--	--	--	--
RDCI	- 0.0187	0.0299	0.531	--	--	--	--
Number of Drugs	- 0.0265	0.0136	0.052	-0.0295	0.0101	0.004	0.97 [0.95 - 0.99]
Number of DMARDs	- 0.0135	0.0708	0.848	--	--	--	--
Marital Status (Single) * Income ($30,000 - $59,999)	- 0.1171	0.2573	0.649	--	--	--	--
Marital Status (Single) * Income ($60,000 or More)	- 0.3179	0.2763	0.250	--	--	--	--
Health Status (Excellent/Good) * Employment (Housework)	- 1.3578	0.3830	< 0.001	--	--	--	--
Health Status (Excellent/Good) * Employment (Paid work)	- 0.7417	0.2995	0.013	--	--	--	--
Health Status (Excellent/Good) * Employment (Retired)	- 0.2174	0.3804	0.568	--	--	--	--
Health Status (Excellent/Good) * Employment (Student)	1.0879	0.7898	0.168	--	--	--	--
Health Status (Excellent/Good) * Employment (Unemployed)	- 0.1175	0.4908	0.811	--	--	--	--

Note. Reference categories for categorical variables are given in Table E1. All variables are assessed as of first survey. SLE = systemic lupus erythematosus. SLE = systemic lupus erythematosus; PCS = physical component scale; RDCI = rheumatic disease comorbidity index; DMARDs = disease-modifying anti-rheumatic drugs.

[a] Type 3 analysis of effect indicated a significant overall effect for these categorical variables (Recruitment: *p* = 0.005; Employment: *p* = 0.029; Income: *p* = 0.015). [b] Confidence interval is inclusive of 1. Odds ratio may be informative but should not be considered definitive.

Table F4

Logistic Regression Model Selection for SLE Group

#	Description	Action Taken	AIC
1	Full Model		2516.41
2	Interim Model	Removed DMARDs	2514.42
3	Interim Model	Removed Marital Status * Income	2511.75
4	Interim Model	Removed Marital Status	2530.56
5	Interim Model	Removed Education	2923.49
6	Interim Model	Removed PCS	3172.43
7	Final Model	Removed Health Status * Employment	3180.58

Note. SLE = systemic lupus erythematosus; DMARDs = disease-modifying anti-rheumatic drugs; PCS = physical component scale.

Table F5

Selected Pairwise Odds Ratios for the SLE Group

Recruitment		**Employment**	
Comparison	**OR [95% CI]**	**Comparison**	**OR [95% CI]**
Provider Referral vs. Self-Enrolled	1.25 [1.03, 1.52]	Disabled vs. Paid Work	1.43 [1.12, 1.82]
Provider Referral vs. Drug Registries	1.36 [0.99, 1.85] [a]	Paid Work vs. Retired	1.01 [0.74, 1.38] [a]
Self-Enrolled vs. Drug Registries	1.09 [0.78, 1.50] [a]	Unemployed vs. Paid Work	1.53 [1.05, 2.23]
Income		Disabled vs. Retired	1.44 [1.05, 1.98]
Comparison	**OR [95% CI]**	Unemployed vs. Retired	1.54 [0.98, 2.43] [a]
Less Than $30,000 vs. $30,000 - $59,999	1.00 [0.81, 1.24] [a]		
Less Than $30,000 vs. $60,000 or More	1.34 [1.06, 1.61]		
$30,000 - $59,999 vs. $60,000 or More	1.31 [1.06, 1.61]		

Note. SLE = systemic lupus erythematosus.

[a] Confidence interval is inclusive of 1. Odds ratio may be informative but should not be considered not definitive.

Appendix G – Logistic Regression Model for Other Rheumatic Diseases Group

Results of logistic regression models for the other rheumatic diseases group are presented in Tables G1, G2, G3, and G4. Refer to Appendix E for more information. HAQ and PCS scores were moderately correlated (*r* = - 0.720, *p* < 0.001). As with the RA group, PCS was dropped in favor of HAQ due to more missing values for PCS (16.7% missing PCS vs. 10.4% missing HAQ). Number of drugs, whose odds ratio was very close to 1, was eliminated due to an excess of qualifying predictors. Model selection proceeded with the remaining predictors. The final model (Model 7) contained an interaction between age and employment whose overall effect was significant (*p* < 0.001) but for which only 1 of the 5 levels (retired patients) was significant on its own. The case was the same with the main effect of employment. A reduced model (Model 8) was fit without the interaction term. Model 8 was a significantly poorer fit (*p* = 0.004) but had only a slightly higher AIC (13548.30 vs. 13531.13). No level of the main effect of unemployment was significant in Model 8 (Type 3 *p* = 0.099), so model selection continued with a removal of this main effect (Model 9). However, the AIC increased considerably to 13816.067. Ultimately Model 7, which included both the employment main effect and the employment-age interaction, was selected as final.

Table G1

Univariable Logistic Regression Results for Other Rheumatic Diseases Group

Parameter	*b*	SE	*p*	OR [95% CI]	Notes
Sex (Male)	0.2114	0.0434	<.001	1.24 [1.13 - 1.35]	Reference group: female.
Education Level (Years)	-0.0516	0.0075	<.001	0.95 [0.94 - 0.96]	OR for unit of 4 years was 0.81 [95% CI 0.77, 0.86].
Race (Black)	0.3253	0.0884	<.001	1.38 [1.16 - 1.65]	Reference group: White.
Race (Other)	0.2880	0.0870	0.001	1.33 [1.12 - 1.58]	
Recruitment (Provider Referral)	-0.0175	0.0683	0.798	0.98 [0.86 - 1.12]	Reference group: other recruitment methods.
Recruitment (Self-Enrolled)	0.0471	0.0728	0.517	1.05 [0.91 - 1.21]	
Recruitment (Drug Registries)	0.1666	0.0856	0.052	1.18 [1.00 - 1.40]	
Age (Years)	-0.0143	0.0012	<.001	0.99 [0.98 - 0.99]	OR for unit of 10 years was 0.87 [95% CI 0.85, 0.89].
Marital Status (Single)	-0.1950	0.0345	<.001	0.82 [0.77 - 0.88]	Reference group: partnered.
Employment (Housework)	-0.4124	0.0636	<.001	0.66 [0.58 - 0.75]	Reference group: disabled.
Employment (Paid)	-0.1901	0.0509	<.001	0.83 [0.75 - 0.91]	
Employment (Retired)	-0.3880	0.0526	<.001	0.68 [0.61 - 0.75]	
Employment (Student)	0.5310	0.1658	0.001	1.70 [1.23 - 2.35]	

Table G1

Univariable Logistic Regression Results for Other Rheumatic Diseases Group

Parameter	*b*	SE	*p*	OR [95% CI]	Notes
Employment (Unemployed)	0.1333	0.0901	0.139	1.14 [0.96 - 1.36]	
Income ($30,000 - $59,999)	-0.1678	0.0436	<.001	0.85 [0.78 - 0.92]	Reference group: income ≤ $30,000.
Income ($60,000 or More)	-0.1464	0.0424	0.001	0.86 [0.79 - 0.94]	
RUCA Category (Urban)	0.0010	0.0371	0.978	1.00 [0.93 - 1.08]	Reference group: rural/isolated.
HAQ	0.1014	0.0256	<.001	1.11 [1.05 - 1.16]	
PCS	-0.0040	0.0017	0.016	1.00 [0.99 - 1.00]	Odds ratio for a change of 10 units was 0.96 [0.93, 0.99]. Dropped from consideration due to significant correlation with HAQ.
MCS	-0.0184	0.0015	<.001	0.98 [0.98 - 0.98]	Odds ratio for a change of 10 units was 0.83 [0.81, 0.86].
RDCI	0.0464	0.0102	<.001	1.05 [1.03 - 1.07]	
Health Status (Excellent/Good)	0.3177	0.0335	<.001	1.37 [1.29 - 1.47]	Reference group: fair/poor.
Number of Drugs	-0.0134	0.00382	<.001	0.99 [0.98 - 0.99]	Dropped from consideration due to odds ratio close to 1.
Number of DMARDs	0.1548	0.0295	<.001	1.17 [1.10 - 1.24]	
Analgesic Use (Yes)	-0.0764	0.0332	0.021	0.93 [0.87 - 0.99]	
Heart Problem (Yes)	0.0470	0.0556	0.398	1.05 [0.94 - 1.17]	
Infection (Yes)	0.2447	0.0342	<.001	1.28 [1.19 - 1.37]	

Note. All variables are assessed as of first survey. RUCA = Rural Urban Commuting Area Codes; HAQ = health assessment questionnaire; PCS = physical component scale; MCS = mental component scale; RDCI = rheumatic disease comorbidity index; DMARDs = disease-modifying anti-rheumatic drugs.

Table G2

Multivariable Logistic Regression Results for Other Rheumatic Diseases Group

	Full Model			Final Model			
Parameter	*b*	SE	*p*	*b*	SE	*p*	OR [95% CI]
Intercept	1.8173	0.3895	< 0.001	1.8741	0.3457	< 0.001	6.51 [3.31, 12.83]
Sex (Male)	0.2087	0.0619	0.001	0.2401	0.0574	< 0.001	1.27 [1.14, 1.42]
Age (Years)	- 0.0150	0.0060	0.012	- 0.0166	0.00565	0.003	0.98 [0.97, 0.99] [a]
Race (Black)	0.3109	0.1117	0.005	0.2702	0.1069	0.012 [b]	1.31 [1.06, 1.62]
Race (Other)	0.2319	0.1075	0.031	0.2117	0.1024	0.039 [b]	1.24 [1.01, 1.51]
Marital Status (Single)	- 0.1364	0.0528	0.010	- 0.0942	0.0459	0.040	0.91 [0.83, 1.00]
Education Level (Years)	- 0.0563	0.0103	< 0.001	- 0.0525	0.0092	< 0.001	0.95 [0.93, 0.97]
Recruitment (Provider Referral)	0.1600	0.0892	0.073	0.1232	0.0827	0.136 [b]	1.13 [0.96, 1.33] [c]
Recruitment (Self-Enrolled)	0.0401	0.0944	0.671	- 0.0169	0.0884	0.849 [b]	0.98 [0.83, 1.17] [c]
Recruitment (Drug Registries)	0.3582	0.1093	0.001	0.2987	0.1029	0.003 [b]	1.35 [1.10, 1.65]
Employment (Housework)	0.1581	0.4281	0.712	- 0.2802	0.4041	0.488 [b]	0.76 [0.34, 1.67] [a, c]
Employment (Paid Work)	0.0348	0.3589	0.923	- 0.0250	0.3422	0.942 [b]	0.98 [0.50, 1.91] [a, c]
Employment (Retired)	- 1.6271	0.4897	0.001	- 1.8007	0.4601	< 0.001 [b]	0.17 [0.07, 0.41] [a]
Employment (Student)	0.7004	0.6200	0.259	0.6493	0.5796	0.263 [b]	1.91 [0.61, 5.96] [a, c]
Employment (Unemployed)	0.4863	0.5737	0.397	0.2621	0.5433	0.630 [b]	1.30 [0.45, 3.77] [a, c]
Income ($30,000 - $59,999)	0.0402	0.0594	0.499	--	--	--	--
Income ($60,000 or More)	0.1047	0.0665	0.115	--	--	--	--
HAQ	- 0.0545	0.0436	0.212	--	--	--	--
MCS	- 0.0125	0.0021	< 0.001	- 0.0114	0.00184	< 0.001	0.99 [0.99, 0.99]
RDCI	0.0133	0.0152	0.382	--	--	--	--

Table G2

Multivariable Logistic Regression Results for Other Rheumatic Diseases Group

	Full Model			Final Model			
Parameter	***b***	**SE**	***p***	***b***	**SE**	***p***	**OR [95% CI]**
Number of DMARDs	0.0501	0.0411	0.223	--	--	--	--
Health Status (Excellent/Good)	0.0504	0.0536	0.348	--	--	--	--
Analgesic Use (Yes)	- 0.0652	0.0456	0.153	--	--	--	--
Infection (Yes)	0.1592	0.0459	0.001	0.1744	0.0432	< 0.001	1.19 [1.09, 1.30]
Age * Employment (Housework)	- 0.0040	0.0076	0.600	0.0039	0.0071	0.587[b]	1.00 [0.99, 1.02][a]
Age * Employment (Paid Work)	- 0.0017	0.0067	0.795	- 0.0003	0.00642	0.959[b]	1.00 [0.99, 1.01][a]
Age * Employment (Retired)	0.0248	0.0079	0.002	0.0273	0.0074	< 0.001[b]	1.03 [1.01, 1.04][a]
Age * Employment (Student)	- 0.0149	0.0160	0.351	- 0.0136	0.0149	0.361[b]	0.99 [0.96, 1.02][a]
Age * Employment (Unemployed)	- 0.0069	0.0110	0.531	- 0.0018	0.0103	0.859[b]	1.00 [0.98, 1.02][a]

Note. Reference categories for categorical variables are given in Table F1. All values are assessed as of the first survey. HAQ = health assessment questionnaire; MCS = mental component score; RDCI = rheumatic disease comorbidity index.

[a] Because the model contains an interaction between age and employment category, the odds ratios given for each main effect and for the interaction terms are valid only in certain cases. See Table 3 for odds ratios interpreted at clinically applicable levels. [b] Type 3 analysis of effect indicated a significant overall effect for the predictor (race: *p* = 0.006; recruitment: *p* < 0.001; employment: *p* < 0.001; age*employment: *p* < 0.001). [c] Confidence interval is inclusive of 1. Odds ratio may be informative but should not be considered definitive.

Table G3

Logistic Regression Model Selection for Other Rheumatic Diseases Group

#	Description	Action Taken	AIC
1	Full Model		12223.38
2	Interim Model	Removed RDCI	12225.74
3	Interim Model	Removed Income	13116.17
4	Interim Model	Removed DMARDs	13116.05
5	Interim Model	Removed HAQ	13191.88
6	Interim Model	Removed Health Status	13529.98
7	Final Model	Removed Analgesic Use	13531.13
8	*Exploratory model*	*Removed Employment * Age*	*13548.30*
9	*Exploratory model*	*Removed Employment*	*13816.07*

Note. RA = rheumatoid arthritis; SLE = systemic lupus erythematosus; RUCA = Rural Urban Commuting Area Codes; HAQ = health assessment questionnaire; PCS = physical component scale; MCS = mental component scale; RDCI = rheumatic disease comorbidity index.

Table G4

Selected Pairwise Odds Ratios for Recruitment Type in the Other Rheumatic Diseases Group

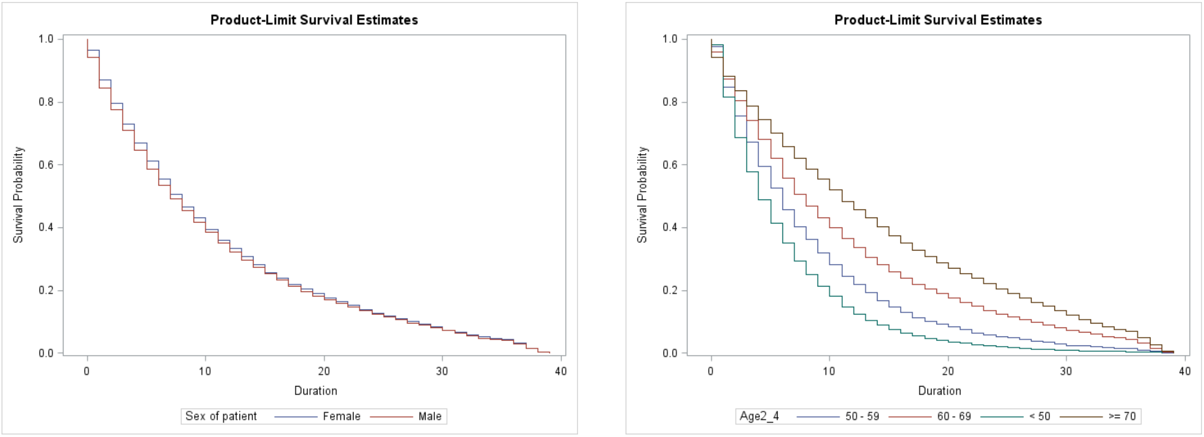
Comparison	OR [95% CI]
Provider Referral vs. Self-Enrolled	1.15 [1.04, 1.27]
Drug Registries vs. Provider Referral	1.19 [1.03, 1.37]
Drug Registries vs. Self-Enrolled	1.37 [1.18, 1.60]

Appendix H – Survival Model for RA Group

Kaplan-Meier survival plots were constructed for each predictor being evaluated for inclusion in the RA group survival model. These included all variables with adequate heterogeneity and a sufficiently low level of missing values. Observations from each patient's first and last surveys were considered, along with difference between the two values if feasible. The Kaplan-Meier plots contained separate curves for each level of the predictor under assessment, with continuous variables categorized into similarly sized groups. Using the LIFETEST procedure in SAS, a chi-square test was performed to identify significant inequality between strata. Predictors with *p* < 0.05 in this test were considered viable candidates for the model. In cases where multiple observations (first, last, or change) for the same variable qualified, only one was permitted. Preference was given to the last survey observation or longitudinal change, and ultimately no first-survey observations were selected for any group. The remaining set of predictors was further reduced by subjective evaluation of differences in the Kaplan-Meier plots and by degree of clinical interest, and from these the full multivariable model was constructed. Figure H1 contains the Kaplan-Meier plots for all variables appearing in the full model.

Figure H1

Kaplan-Meier Survival Plots by Selected Predictors for RA Group



Sex

Age

Figure H1

Kaplan-Meier Survival Plots by Selected Predictors for RA Group

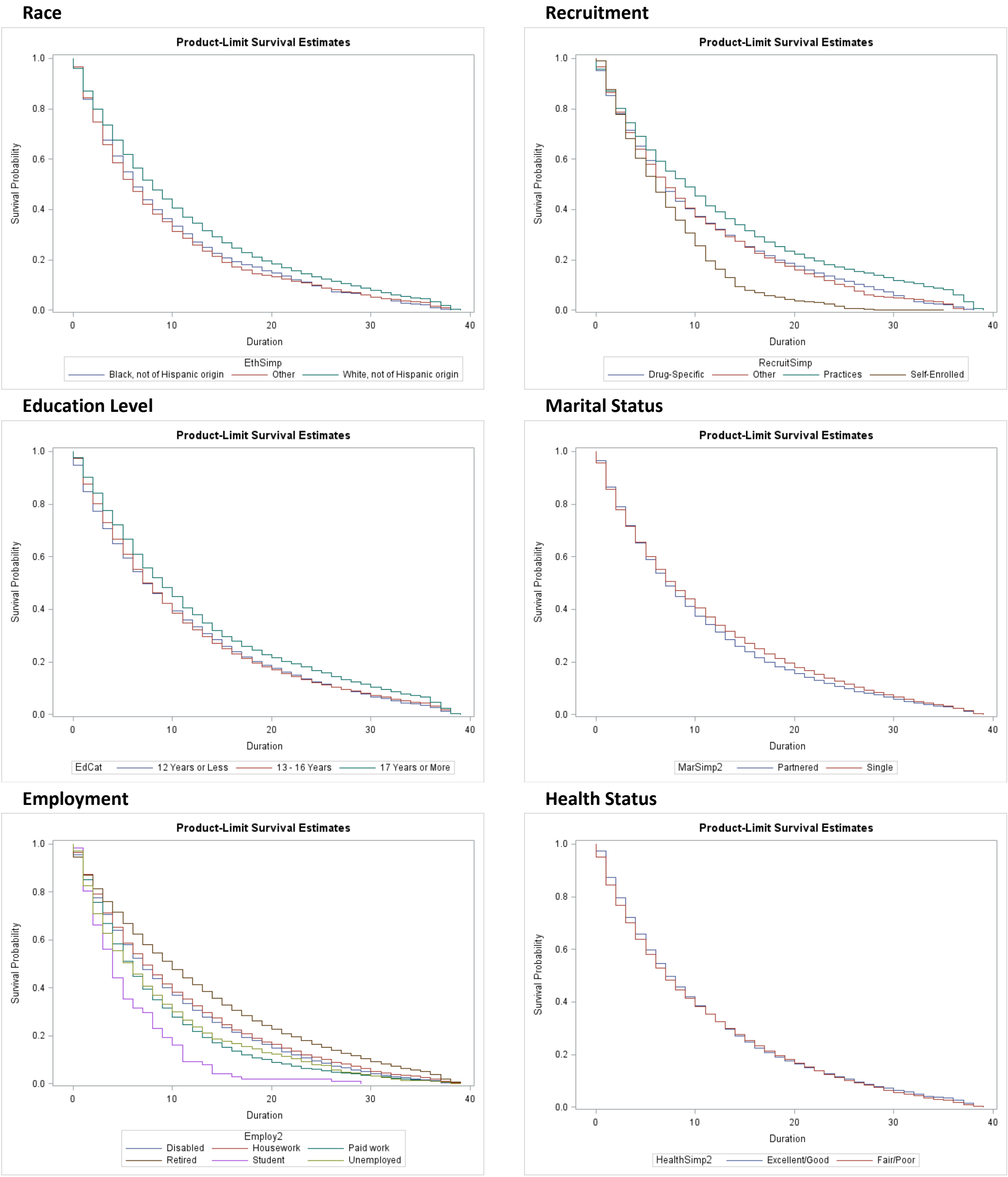
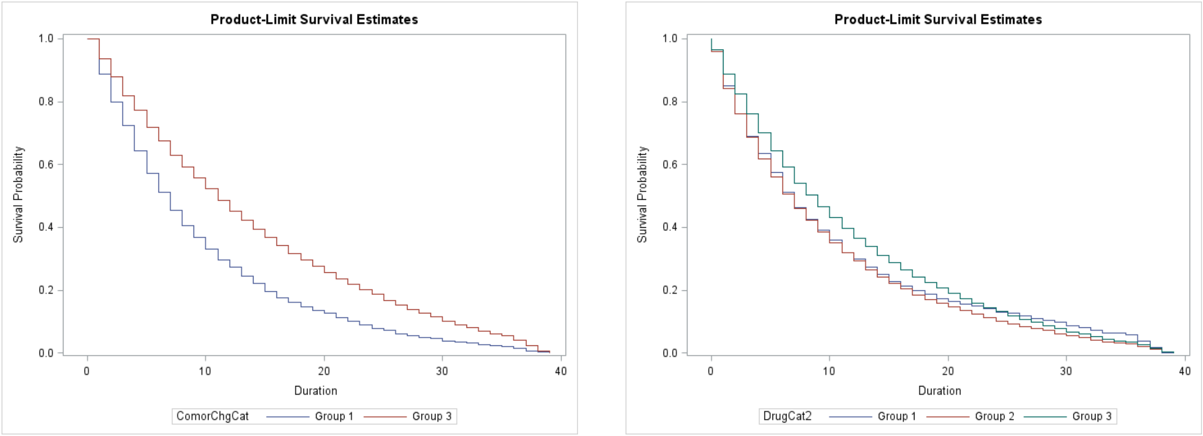


Figure H1

Kaplan-Meier Survival Plots by Selected Predictors for RA Group



Change in RDCI

Number of Drugs [a]

Note. RA = rheumatoid arthritis; RDCI = rheumatic disease comorbidity index.

The proportional hazards assumption for the survival models was evaluated using Schoenfeld residuals and observed vs. expected plots. A summary of findings is presented in Table H1. A *p*-value of less than 0.05 in the Schoenfeld residuals test indicated that residuals for that variable were significantly correlated with duration. This was suggestive of failure to meet the proportional hazards assumption. In the observed vs. expected plots, shown in Figure H2, the actual survival probability curve for each level of the variable under evaluation is compared to a predicted curve. As with the Kaplan-Meier plots, continuous variables have been categorized. Blatant inconsistency between corresponding observed and expected curves is suggestive of failure to meet the proportional hazards assumption.

In the case of the recruitment variable, both the Schoenfeld residuals test and the observed vs. expected plot indicated that failure to meet the proportional hazards assumption, and as a result the model was stratified on this variable. The Schoenfeld test also suggested that age, marital status, and health status violated the assumption ($p < 0.001$ in each case); however, the observed vs. expected plots for each predictor were highly consistent. Conversely, the observed vs. expected plots for employment indicated a possible violation, but a *p*-value of 0.688 in the Schoenfeld test suggested otherwise. Each of these variables was retained as a predictor in the full model without stratification.

Table H1

Evaluation of Proportional Hazards Assumption for RA Group

Parameter	Schoenfeld Residuals (*p*)	Observed vs. Expected Plots	Notes
Sex	0.097	Highly Consistent	
Education Level	0.277	Highly Consistent	Education level was categorized as ≤ 12 years, 13 – 16 years, and ≥ 17 years.
Race	0.216	Highly Consistent	
Recruitment	0.012 [a]	Not Consistent [a]	Selected for stratification.
Age	< 0.001 [a]	Highly Consistent	Age was categorized as < 50 years, 51 – 60 years, 61 – 70 years, and ≥ 70 years.
Marital Status	< 0.001 [a]	Highly Consistent	
Employment	0.688	Not Consistent [a]	
Health Status	< 0.001 [a]	Highly Consistent	
Number of Drugs	0.742	Acceptable	Number of drugs was categorized as ≤ 4, 5 – 7, and ≥ 8.
Change in RDCI	0.375	Highly Consistent	Change in RDCI was categorized as decreased, remained the same, and increased.

Note. All dynamic variables are assessed as of the last survey. RA = rheumatoid arthritis; RDCI = rheumatic disease comorbidity index.

[a] Test suggests failure to meet the proportional hazards assumption.

Figure H2

Observed vs. Expected Plots by Selected Predictors for RA Group

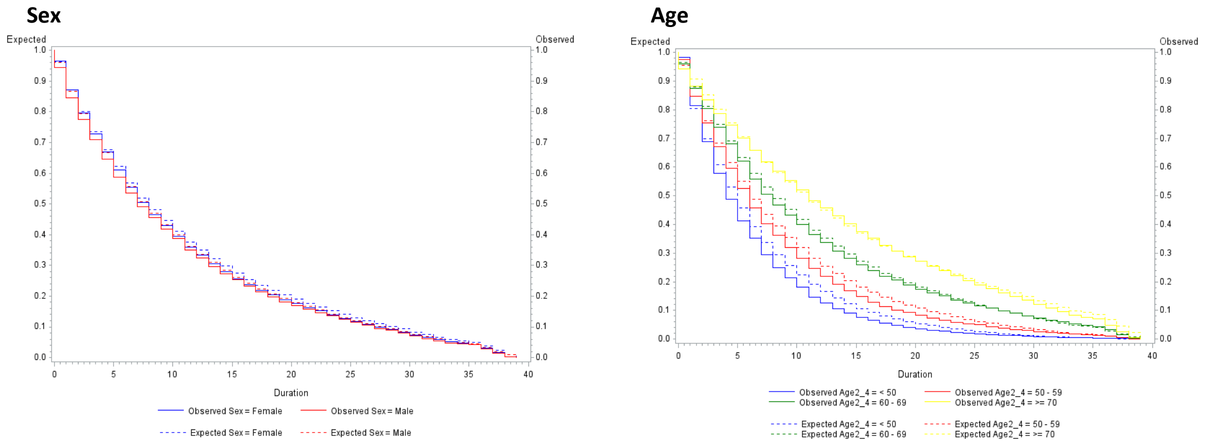
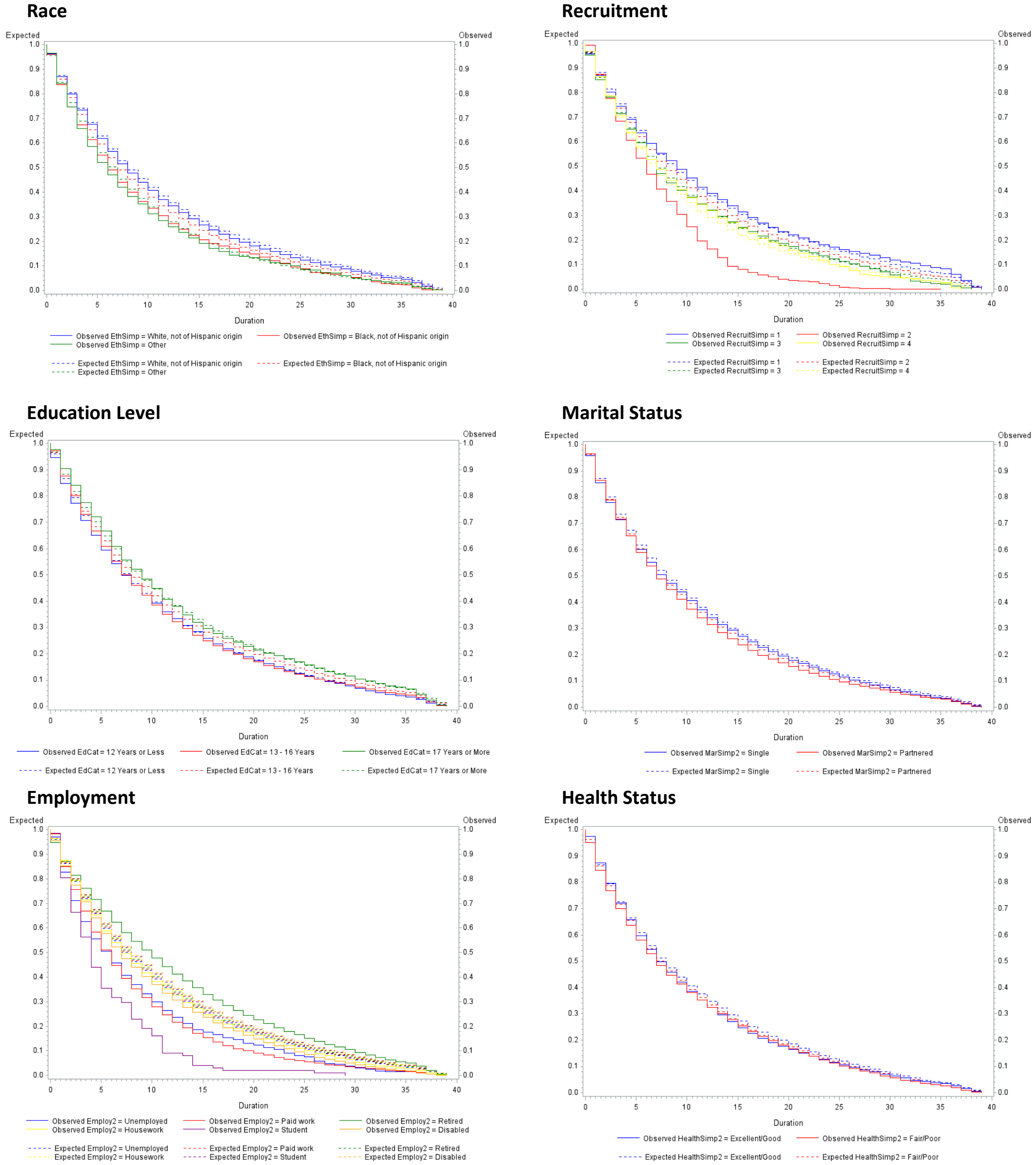


Figure H2

Observed vs. Expected Plots by Selected Predictors for RA Group



Race

Recruitment

Education Level

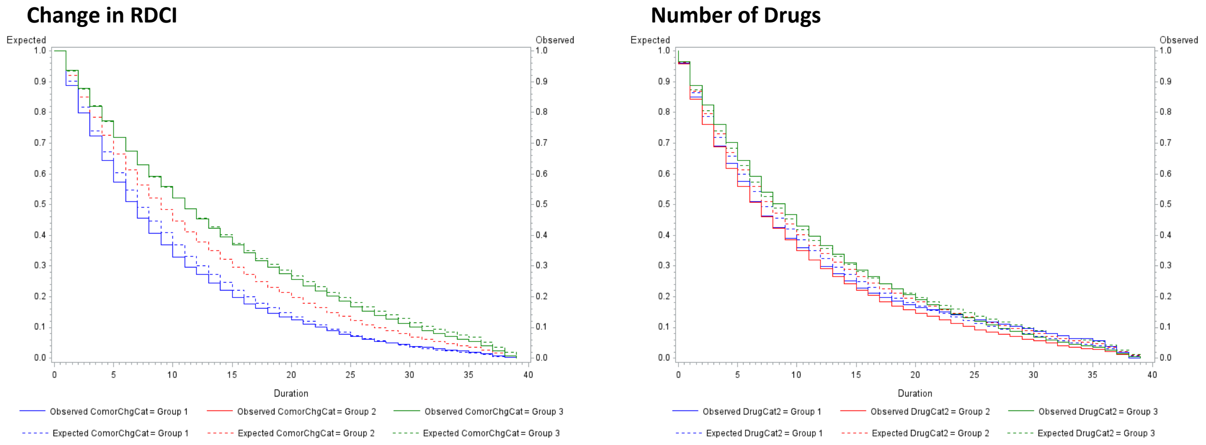
Marital Status

Employment

Health Status

Figure H2

Observed vs. Expected Plots by Selected Predictors for RA Group


Change in RDCI

Number of Drugs

Note. RA = rheumatoid arthritis; RDCI = rheumatic disease comorbidity index.

The full model, stratified on recruitment, was considered both without interaction (Model 1) and with interaction (Model 2). Full specification of Model 1 is given in Table H2, and selected information about Model 2 is given in Table H3. Model 2 performed slightly better than Model 1 (total AIC 267,662.10 vs. 267,678.06), but the difference was not statistically significant (*p* = 0.316). Model 1 contained only a single non-significant term (the indicator variable for other race); however, Type 3 analysis of effect indicated that the overall effect of the predictor was significant (*p* = 0.001), and the predictor was retained. Model 1 was selected as the final model.

Table H2

Survival Model for RA Group Stratified on Recruitment Without Interaction

Parameter	*b*	SE	*p*	Hazard Ratio [95% CI]	Notes
Sex (Male)	0.1482	0.0198	< 0.001	1.16 [1.12, 1.21]	Reference group: Female.
Age (Years)	- 0.0264	0.0007	< 0.001	0.97 [0.97, 0.98]	
Race (Black)	0.1331	0.0369	< 0.001	1.14 [1.06, 1.23]	Reference group: White. Odds ratio for other race vs. White was inconclusive due to the insignificant *p*-value for the parameter estimate.
Race (Other)	0.0268	0.0326	0.411	1.03 [0.96, 1.09]	
Marital Status (Single)	- 0.0498	0.0164	0.002	0.95 [0.92, 0.98]	Reference group: Partnered.
Education Level (Years)	- 0.0460	0.0033	< 0.001	0.96 [0.95, 0.96]	
Employment (Housework)	0.1986	0.0295	< 0.001	1.22 [1.15, 1.29]	Reference group: Disabled.
Employment (Paid work)	0.1150	0.0252	< 0.001	1.12 [1.07, 1.18]	
Employment (Retired)	0.0971	0.0250	< 0.001	1.10 [1.05, 1.16]	
Employment (Student)	0.2208	0.1051	0.036	1.25 [1.01, 1.53]	
Employment (Unemployed)	0.2350	0.0483	< 0.001	1.26 [1.15, 1.39]	
Change in RDCI	- 0.0436	0.0048	< 0.001	0.96 [0.95, 0.97]	

Table H2

Survival Model for RA Group Stratified on Recruitment Without Interaction

Parameter	*b*	SE	*p*	Hazard Ratio [95% CI]	Notes
Health Status (Excellent/Good)	0.0484	0.0161	0.003	1.05 [1.02, 1.08]	Reference group: Fair/Poor.
Number of Drugs	- 0.0144	0.0020	< 0.001	0.99 [0.98, 0.99]	

Note. RA = rheumatoid arthritis; RDCI = rheumatic disease comorbidity index.

Table H3

Survival Model for RA Group Stratified on Recruitment With Interaction

	Provider Referral		Self-Enrolled		Drug Registries		Other	
Parameter	*b*	*p*	*b*	*p*	*b*	*p*	*b*	p
Sex	0.1721	< 0.001	0.2588	< 0.001	0.0942	0.006	0.0916	0.082
Education Level (Years)	- 0.0535	< 0.001	-0.0152	0.086	- 0.0409	< 0.001	- 0.0644	< 0.001
Race (Black)	0.1174	0.045	0.0764	0.543	0.1859	0.002	0.0192	0.859
Race (Other)	- 0.0042	0.943	0.0507	0.533	0.0249	0.659	0.0560	0.459
Age (Years)	- 0.0278	< 0.001	- 0.0239	< 0.001	- 0.0267	< 0.001	- 0.0243	< 0.001
Marital Status (Single)	- 0.0631	0.015	- 0.0852	0.068	- 0.0315	0.267	- 0.0326	0.461
Employment (Housework)	0.2198	< 0.001	0.3901	< 0.001	0.1753	< 0.001	0.0280	0.735
Employment (Paid work)	0.0644	0.122	0.1807	0.006	0.1669	< 0.001	0.0942	0.168
Employment (Retired)	0.1270	0.002	0.1510	0.045	0.0610	0.141	0.0358	0.594
Employment (Student)	0.1900	0.363	0.4110	0.039	0.2637	0.159	- 0.0435	0.876
Employment (Unemployed)	0.0733	0.361	0.4670	< 0.001	0.3057	< 0.001	0.2738	0.034
Change in RDCI	- 0.0575	< 0.001	0.000024	0.999	- 0.0527	< 0.001	- 0.0251	0.053
Health Status (Excellent/Good)	0.0508	0.042	0.0948	0.045	0.0695	0.013	- 0.0302	0.490
Number of Drugs	- 0.0203	< 0.001	- 0.0148	0.007	- 0.0138	< 0.001	- 0.0006	0.908

Note. RA = rheumatoid arthritis; RDCI = rheumatic disease comorbidity index.

As with the logistic regression, comparisons among various levels of the employment variable were of interest. This variable was highly significant, and meaningful odds ratios were discernable in several pairings of levels. Selected comparisons are presented in Table H4, all in the direction resulting in an odds ratio greater than 1.

Table H4

Selected Pairwise Hazard Ratios for Employment in the RA Group

Comparison	OR [95% CI]	Comparison	OR [95% CI]
Paid Work vs. Retired	1.02 [0.97, 1.07] [a]	Retired vs. Disabled	1.10 [1.05, 1.16]
Paid Work vs. Disabled	1.12 [1.07, 1.18]	Housework vs. Disabled	1.22 [1.15, 1.29]
Housework vs. Paid Work	1.09 [1.03, 1.15]	Unemployed vs. Disabled	1.26 [1.15, 1.39]
Unemployed vs. Paid Work	1.13 [1.03, 1.24]	Housework vs. Retired	1.11 [1.05, 1.16]

Note. RA = rheumatoid arthritis.

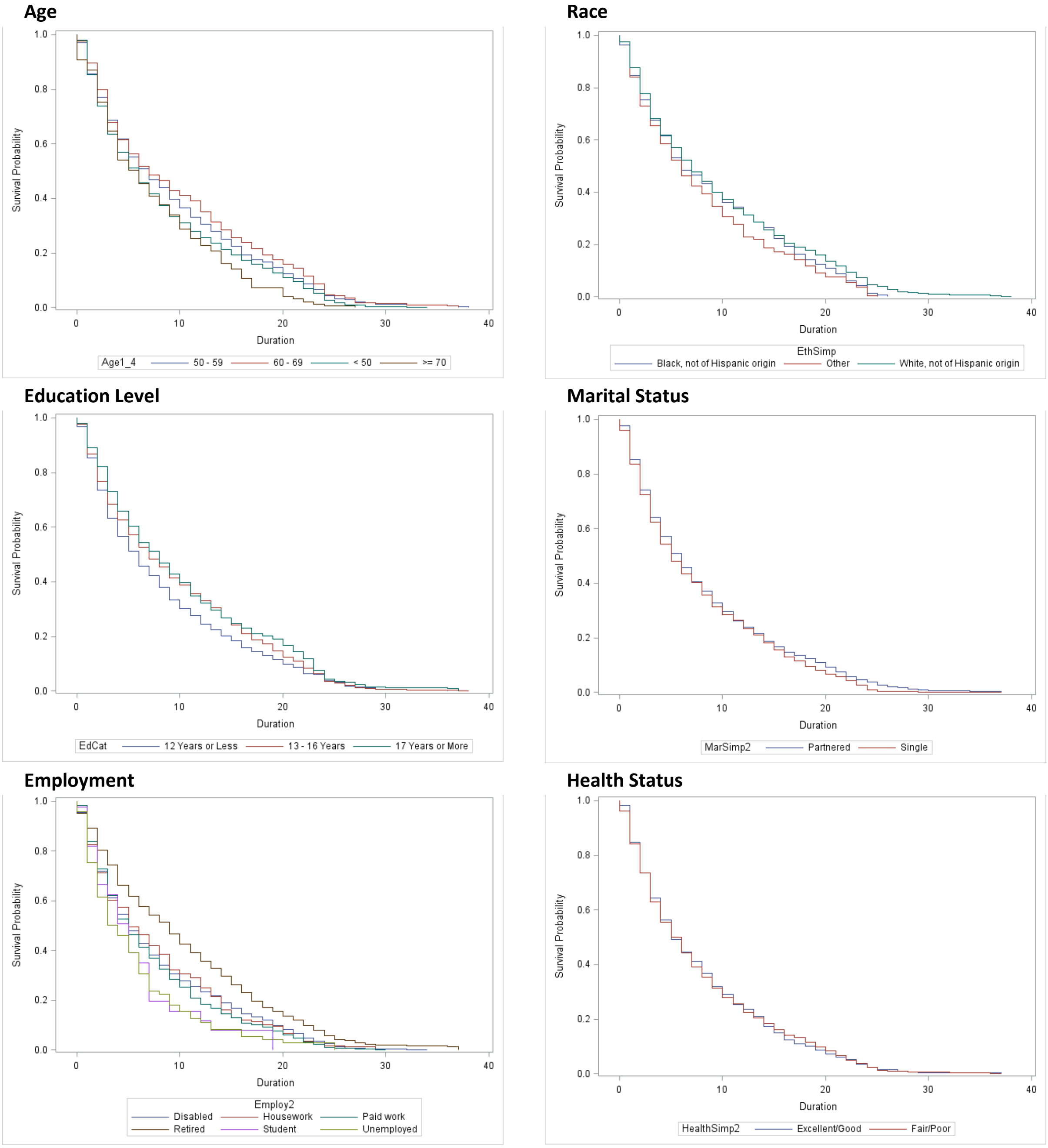
[a] Confidence interval is inclusive of 1. Odds ratio may be informative but should not be considered definitive.

Appendix I – Survival Model for SLE Group

Kaplan-Meier plots for the SLE group are presented in Figure I1. Refer to Appendix H for more information. Plots are shown for all variables included in the full survival model. Dynamic variables are assessed as of the last survey.

Figure I1

Kaplan-Meier Survival Plots by Selected Predictors for SLE Group



Age

Race

Education Level

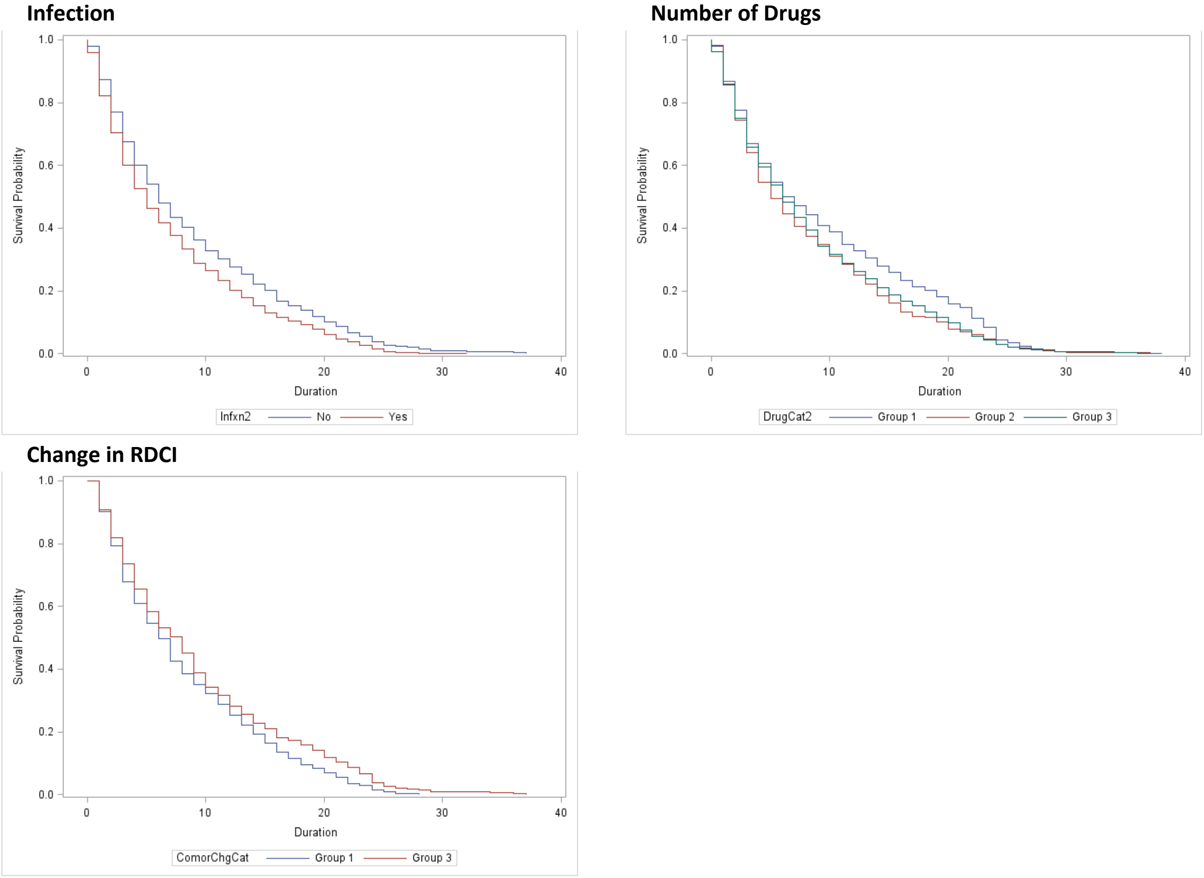
Marital Status

Employment

Health Status

Figure I1

Kaplan-Meier Survival Plots by Selected Predictors for SLE Group



Infection

Number of Drugs

Change in RDCI

Note. SLE = systemic lupus erythematosus; RDCI = rheumatic disease comorbidity index.

Summarized findings for the proportional hazards assumption are given in Table I1, followed by observed vs. expected plots in Figure I2. These processes are described in Appendix H. The Schoenfeld residuals test suggested that marital status violated the assumption (*p* = 0.001); however, the observed vs. expected plots were highly consistent. The opposite occurred with the employment variable (Schoenfeld *p* = 0.296). Each of these variables was retained as a predictor in the model, and no variables were removed for stratification.

Table I1

Evaluation of Proportional Hazards Assumption for SLE Group

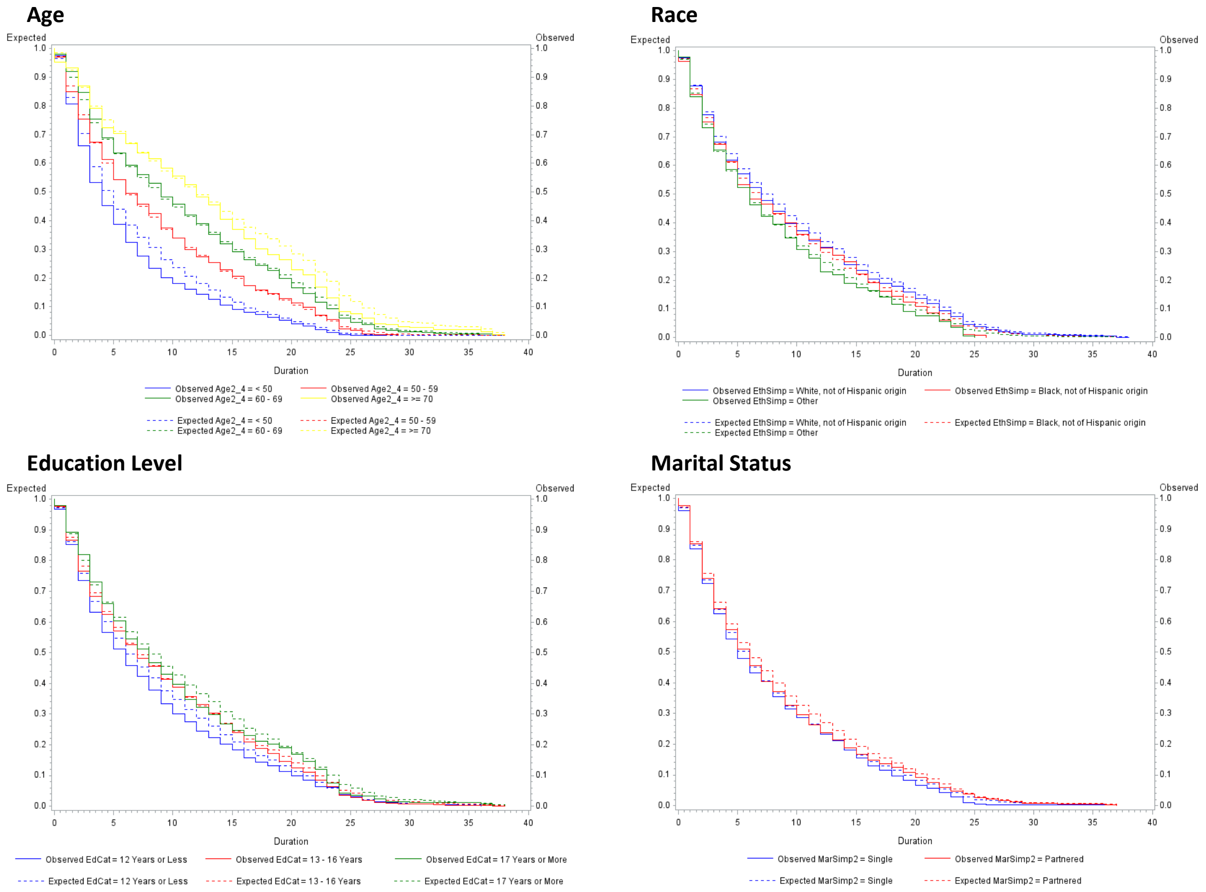
Parameter	Schoenfeld Residuals (*p*)	Observed vs. Expected Plots	Notes
Age	0.845	Highly Consistent	Categorized as < 50 years, 51 – 60 years, 61 – 70 years, and ≥ 70 years.
Race	0.472	Highly Consistent	
Education Level	0.170	Acceptable	Categorized as ≤ 12 years, 13 – 16 years, and ≥ 17 years.
Marital Status	0.001 [a]	Highly Consistent	
Employment	0.296	Not Consistent [a]	
Infection	0.609	Highly Consistent	
Number of Drugs	0.188	Acceptable	Categorized as ≤ 4, 5 – 7, and > 7.
Change in HAQ II Score	0.840	Acceptable	Categorized as < - 0.1, - 0.1 – 0.1, > 0.1.
Change in RDCI	0.473	Acceptable	Categorized as decreased, remained the same, and increased.

Note. All variables are assessed as of the last survey. SLE = systemic lupus erythematosus; HAQ = health assessment questionnaire; RDCI = rheumatic disease comorbidity index.

[a] Test suggests failure to meet the proportional hazards assumption.

Figure I2

Observed vs. Expected Plots by Selected Predictors for SLE Group



Age

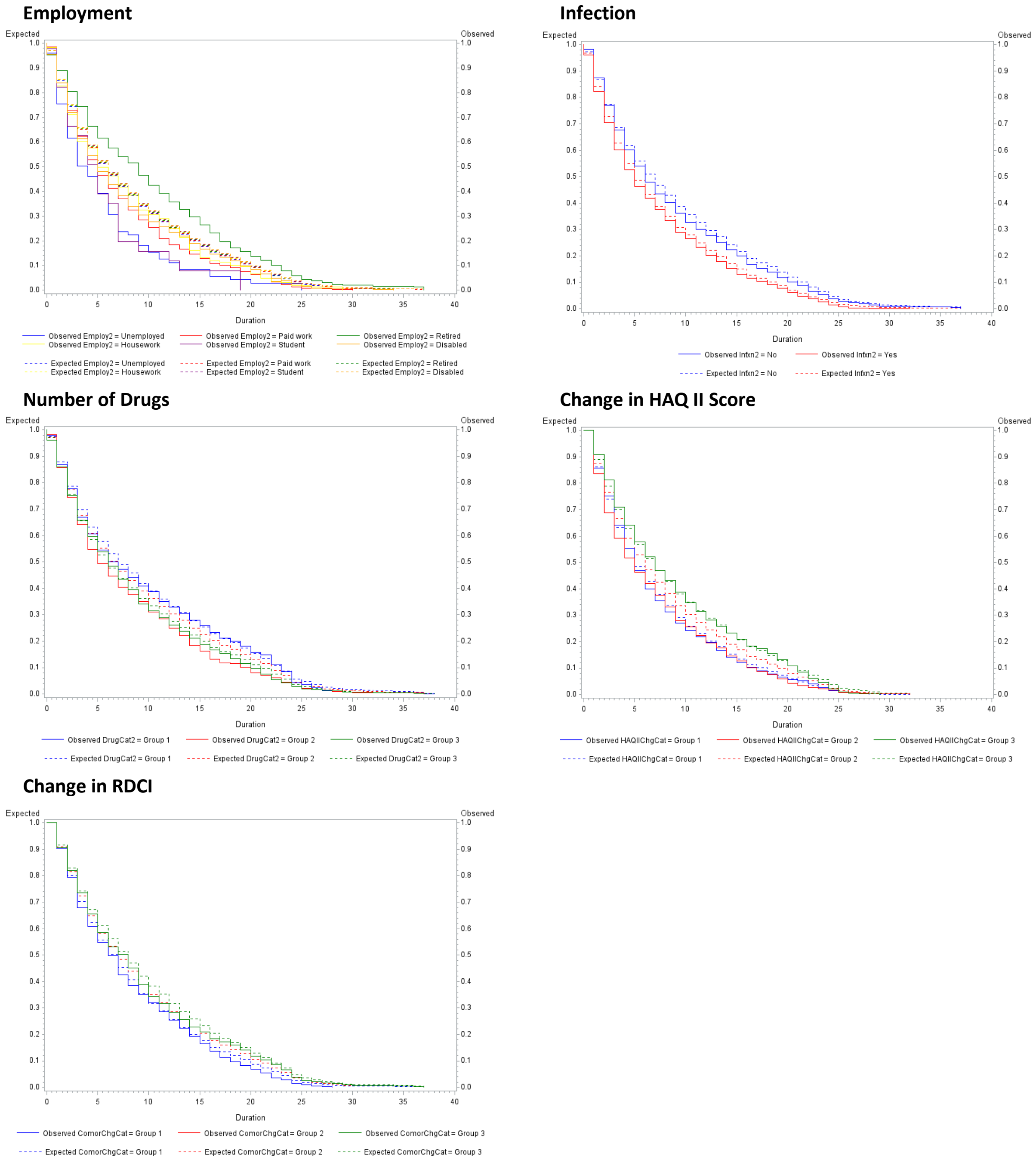
Race

Education Level

Marital Status

Figure I2

Observed vs. Expected Plots by Selected Predictors for SLE Group



Note. SLE = systemic lupus erythematosus; HAQ = health assessment questionnaire; RDCI = rheumatic disease comorbidity index.

The full model (Model 1) contained several predictors whose effects were not significant. These were removed sequentially to arrive at the final model (Model 6). Specification of both Model 1 and Model 6 is presented in Table I2, and the series of iterations is listed in Table I3.

Table I2

Survival Models for SLE Group

	Full Model			Final Model			
Parameter	***b***	**SE**	***p***	***b***	**SE**	***p***	**OR [95% CI]**
Age (Years)	-0.0246	0.0027	< 0.001	-0.0240	0.0021	< 0.001	0.98 [0.97, 0.98]
Race (Black)	-0.0826	0.0837	0.324	--	--	--	--
Race (Other)	-0.0917	0.0932	0.325	--	--	--	--
Marital Status (Single)	-0.0941	0.0598	0.116	--	--	--	--
Education Level (Years)	-0.0468	0.0109	< 0.001	-0.0421	0.0103	< 0.001	0.96 [0.94, 0.98]
Employment (Housework)	-0.0097	0.1046	0.927	--	--	--	--
Employment (Paid work)	-0.0460	0.0756	0.543	--	--	--	--
Employment (Retired)	-0.0324	0.0950	0.733	--	--	--	--
Employment (Student)	-0.1010	0.2290	0.659	--	--	--	--
Employment (Unemployed)	0.1922	0.1468	0.190	--	--	--	--
Infection (Yes)	0.1963	0.0564	0.001	0.1845	0.0534	0.001	1.20 [1.08, 1.34]
Number of Drugs	-0.0054	0.0063	0.391	--	--	--	--
Change in HAQ II Score	-0.1708	0.0547	0.002	-0.1644	0.0504	0.001	0.85 [0.77, 0.94]
Change in RDCI	0.0070	0.0150	0.643	--	--	--	--

Note. All dynamic variables are assessed as of the last survey. SLE = systemic lupus erythematosus; HAQ = health assessment questionnaire; RDCI = rheumatic disease comorbidity index.

Table I3

Survival Model Selection for SLE Group

#	Description	Action Taken	AIC
1	Full Model		16360.867
2	Interim Model	Removed Employment	16353.731
3	Interim Model	Removed Change in RDCI	16351.912
4	Interim Model	Removed Number of Drugs	16350.524
5	Interim Model	Removed Race	16348.171
6	Final Model	Removed Marital Status	17903.622

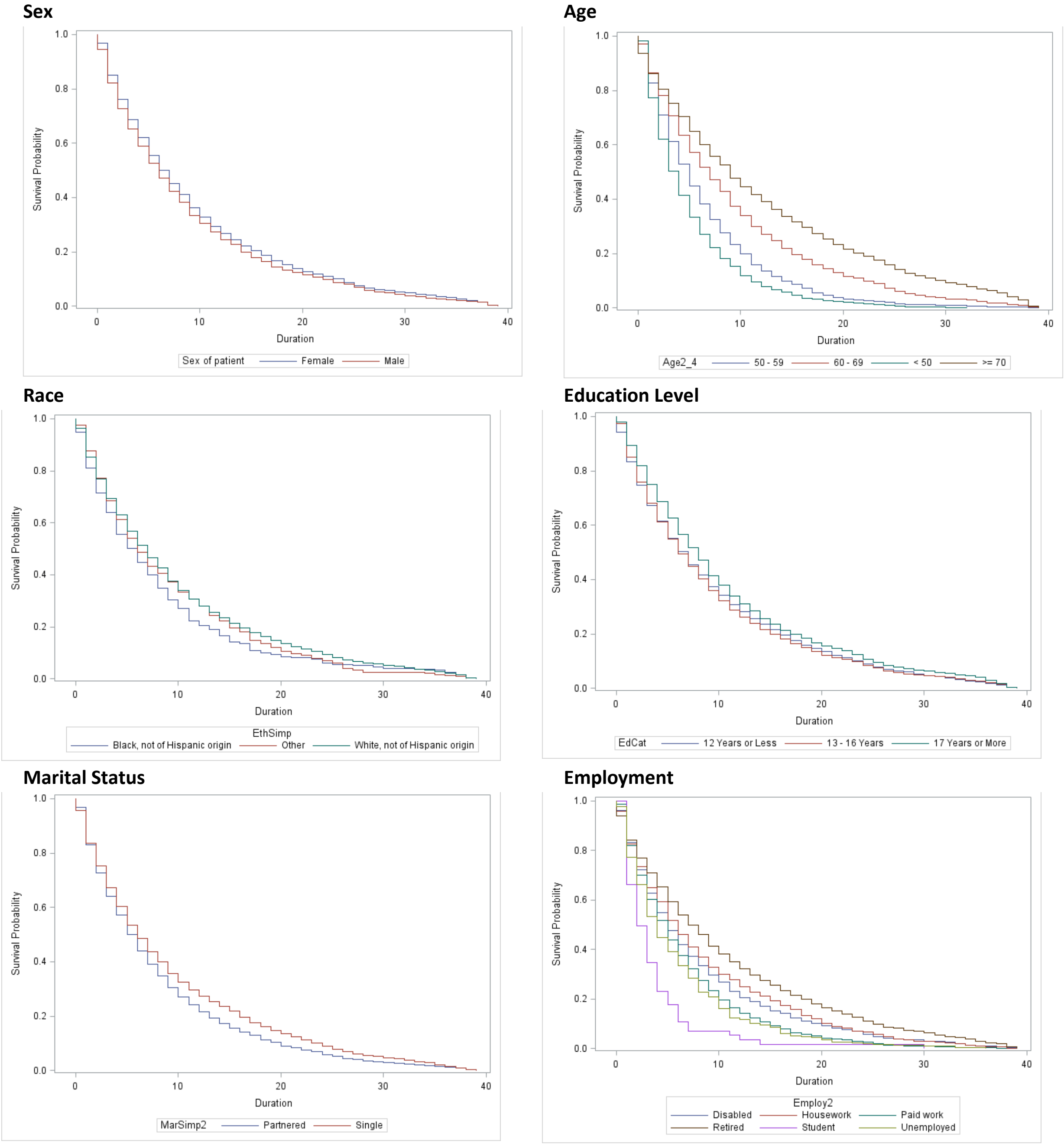
Note. SLE = systemic lupus erythematosus; RDCI = rheumatic disease comorbidity index.

Appendix J – Survival Model for Other Rheumatic Diseases Group

Kaplan-Meier plots for the other rheumatic diseases group are presented in Figure J1. Refer to Appendix H for more information. Plots are shown for all variables included in the full survival model. All dynamic variables are assessed as of the last survey.

Figure J1

Kaplan-Meier Survival Plots by Selected Predictors for Other Rheumatic Diseases Group



Sex

Age

Race

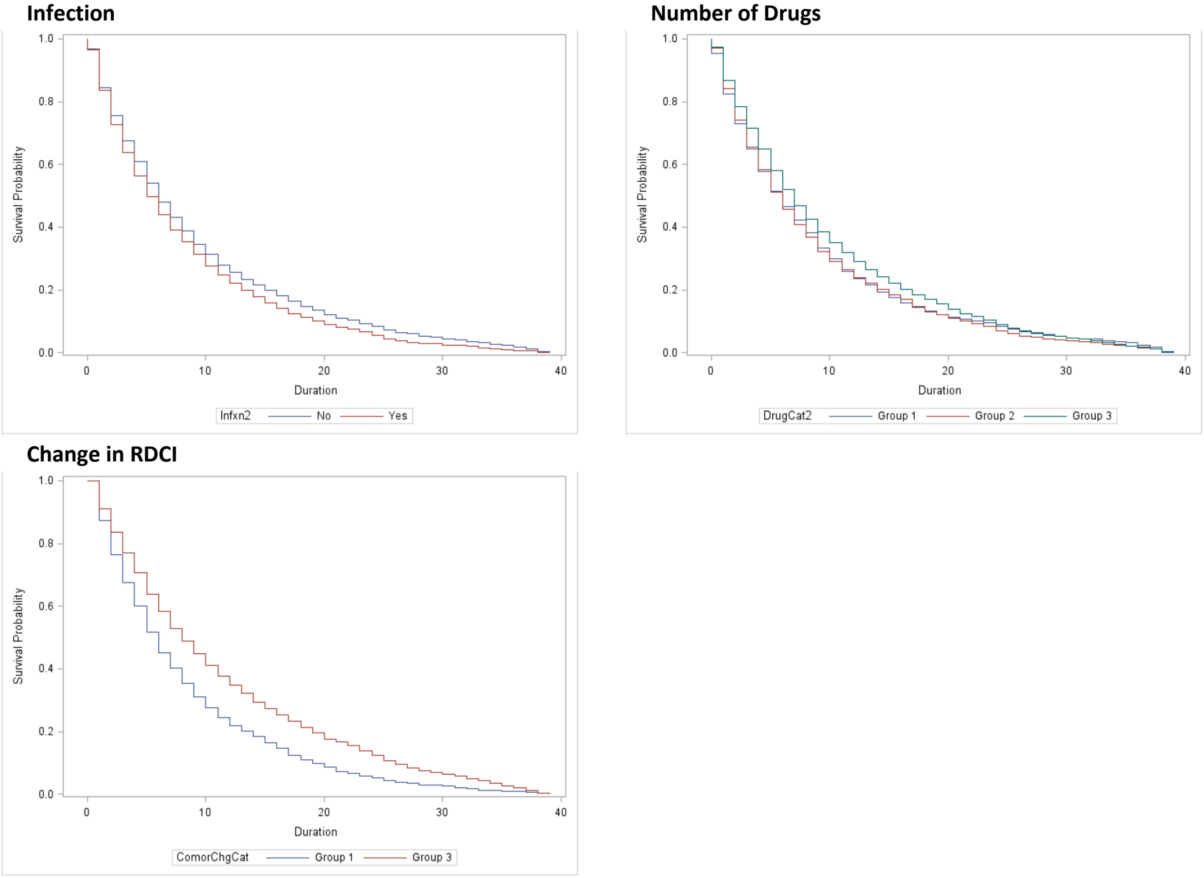
Education Level

Marital Status

Employment

Figure J1

Kaplan-Meier Survival Plots by Selected Predictors for Other Rheumatic Diseases Group



Note. RDCI = rheumatic disease comorbidity index.

Findings for evaluation of the proportional hazards assumption are given in Table J1, and observed vs. expected plots are shown in Figure J2. These processes are described in Appendix H. The Schoenfeld residuals test suggested that age and marital status violated the assumption ($p < 0.001$ for each); however, the observed vs. expected plots for each predictor were highly consistent. Both variables were retained in the model without stratification. Conversely, the observed vs. expected plots for recruitment and employment indicated a possible violation, but the Schoenfeld residuals test suggested otherwise (recruitment $p = 0.237$; employment $p = 0.332$). The inconsistencies in the observed vs. expected plots were highly egregious in both cases, and as a result the model was stratified on both variables.

Table J1

Evaluation of Proportional Hazards Assumption for Other Rheumatic Diseases Group

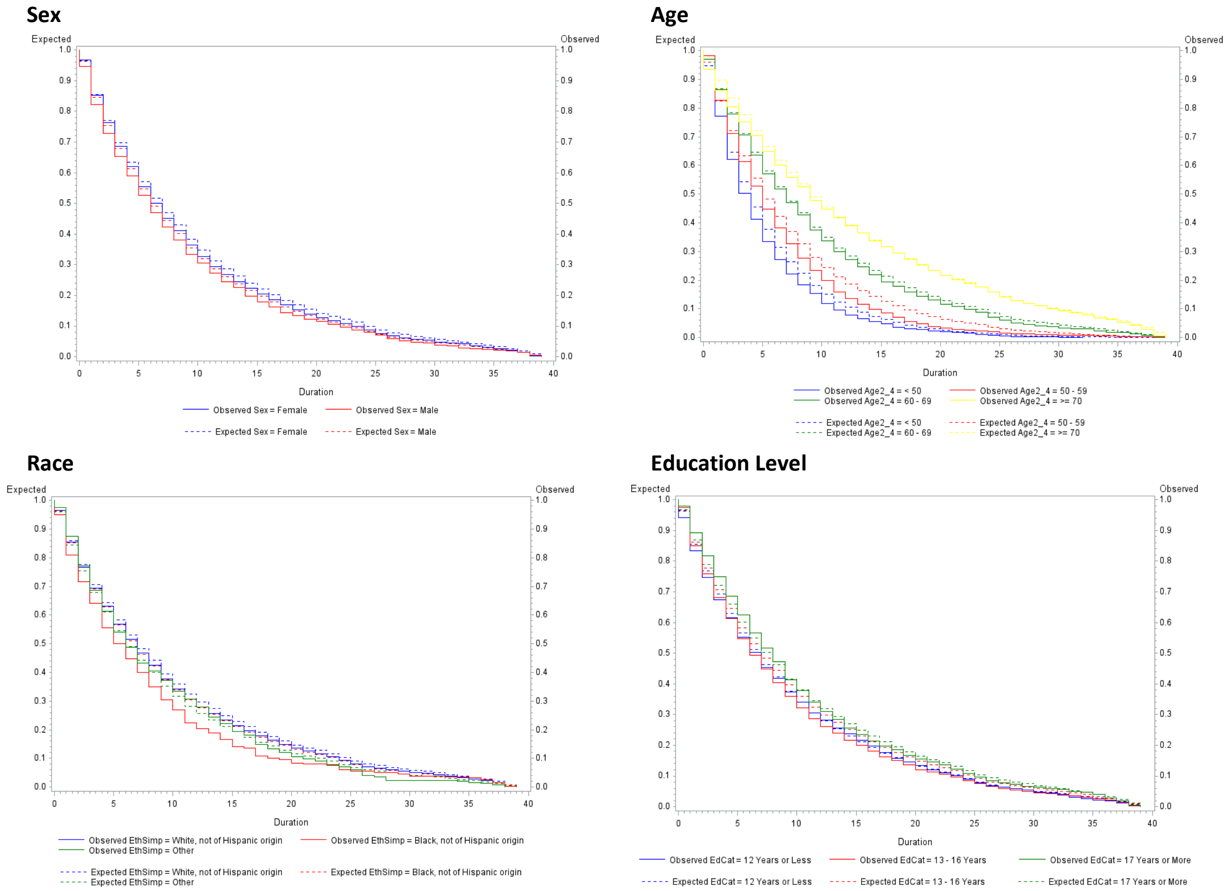
Parameter	Schoenfeld Residuals (*p*)	Observed vs. Expected Plots	Notes
Sex	0.125	Highly Consistent	
Age	< 0.001 [a]	Highly Consistent	Categorized as < 50 years, 51 – 60 years, 61 – 70 years, and ≥ 70 years.
Race	0.238	Acceptable	
Education Level	0.679	Highly Consistent	Categorized as ≤ 12 years, 13 – 16 years, and ≥ 17 years.
Recruitment	0.237	Not Consistent [b]	Selected for stratification.
Marital Status	< 0.001 [a]	Highly Consistent	
Employment	0.332	Not Consistent [b]	Selected for stratification.
Infection	0.845	Highly Consistent	
Number of Drugs	0.112	Highly Consistent	Categorized as ≤ 4, 5 – 7, and ≥ 8.
Change in RDCI	0.077	Highly Consistent	Categorized as decreased, remained the same, and increased.

Note. All dynamic variables are assessed as of the last survey. RDCI = rheumatic disease comorbidity index.

[a] Test suggests failure to meet the proportional hazards assumption.

Figure J2

Observed vs. Expected Plots by Selected Predictors for SLE Group



Sex

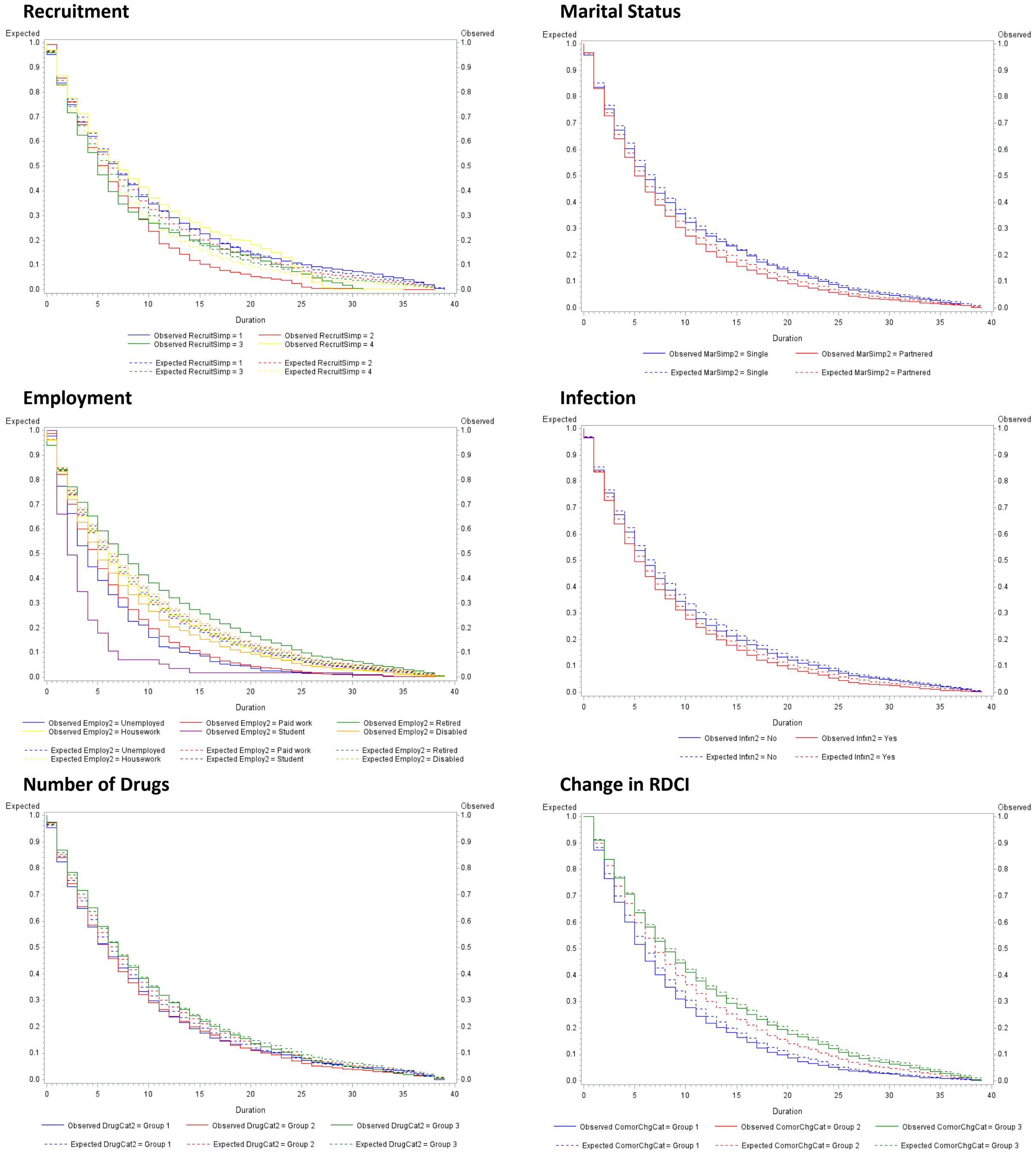
Age

Race

Education Level

Figure J2

Observed vs. Expected Plots by Selected Predictors for SLE Group



Note: SLE = systemic lupus erythematosus; RDCI = rheumatic disease comorbidity index.

The full model, stratified on recruitment and employment, was considered both without interaction (Model 1) and with interaction (Model 2). The interaction model, being stratified on one variable with 4 levels and another with 6 levels, contained 24 sets of parameters. Model 1 performed better as assessed by AIC ($p < 0.001$). Selection proceeded with sequential removal of non-significant predictors from Model 1 to arrive at the final model (Model 5). Specification of both Model 1 and Model 5 is presented in Table J2, and the series of iterations is listed in Table J3. Details of Model 2 are not provided in the interest of brevity.

Table J2

Survival Model for Other Rheumatic Diseases Group Stratified on Recruitment and Employment Without Interaction

	Full Model			**Final Model**			
Parameter	***b***	**SE**	***p***	***b***	**SE**	***p***	**OR [95% CI]**
Sex (Male)	0.1092	0.0333	0.001	0.1071	0.0317	0.001	1.11 [1.05, 1.18]
Age (Years)	-0.0272	0.0012	< 0.001	-0.0265	0.0011	< 0.001	0.97 [0.97, 0.98]
Race (Black)	0.0273	0.0627	0.663	--	--	--	--
Race (Other)	-0.0700	0.0611	0.251	--	--	--	--
Education Level (Years)	-0.0200	0.0051	< 0.001	-0.0213	0.0050	< 0.001	0.98 [0.97, 0.99]
Marital Status (Single)	0.0320	0.0250	0.201	--	--	--	--
Infection (Yes)	0.0264	0.0249	0.289	--	--	--	--
Number of Drugs	-0.0130	0.0031	< 0.001	-0.0150	0.0029	< 0.001	0.99 [0.98, 0.99]
Change in RDCI	-0.0138	0.0072	0.057	-0.0142	0.0070	0.042	0.99 [0.97, 1.00]

Note. All dynamic variables are assessed as of the last survey. RDCI = rheumatic disease comorbidity index.

Table J3

Survival Model Selection for Other Rheumatic Diseases Group

#	**Description**	**Action Taken**	**AIC**
1	Full Model Without Interaction		82798.90
2	Full Model With Interaction		86120.56 [a]
3	Interim No-Interaction Model	Removed Race	83359.57
4	Interim No-Interaction Model	Removed Infection	86621.55
6	Final No-Interaction Model	Removed Marital Status	89482.04

[a] Sum of AIC values for the 24 models that resulted from stratification on recruitment and employment with interaction.