XVII Congreso Galego de Estatística e Investigación de Operacións
Ourense, 23 - 25 de outubro de 2025

Modelling long-term remission in Rheumatoid Arthritis: Insights from a mixture cure model perspective using MDCcure package

Blanca E. Monroy-Castillo[1], M. Amalia Jácome[2] and Ricardo Cao[1]

[1]MODES Research group, CITIC (Center for Information and Communications Technology Research), Department of Mathematics, Faculty of Informatics, Universidade da Coruña, 15071 Campus de Elviña s/n, A Coruña, Spain

[2]MODES Research group, CITIC (Center for Information and Communications Technology Research), Department of Mathematics, Faculty of Science, Universidade da Coruña, 15071 Campus de Elviña s/n, A Coruña, Spain

ABSTRACT

Rheumatoid arthritis (RA) is the most common systemic inflammatory disease, characterized by recurrent flares that contribute to joint damage, disability, and reduced quality of life. While treatment is essential for controlling symptoms and slowing disease progression, it is often associated with substantial side effects. Recent advances suggest that sustained deep remission, including immunological and drug-free remission, may be achievable, raising the question of whether some patients can be considered "cured". This clinical context allows for the application of statistical methodologies such as mixture cure models, which explicitly account for a cured subgroup. In this work, we explore mixture cure modeling to gain insights into long-term remission in RA. Using data from a clinical study with control and experimental groups, we investigate covariates influencing the probability of remission. Analyses were conducted with the `MDCcure` package in R, providing a practical framework for modeling cure rates in RA.

Keywords: R package; nonparametric test; cure probability;

1. INTRODUCTION

Rheumatoid arthritis (RA) is the most frequently diagnosed systemic inflammatory disease, characterized by recurrent flares that contribute to joint damage, disability, and impaired quality of life (Wasserman, 2011; Myasoedova et al., 2019). Diagnosis is based on clinical features such as swelling in at least one joint, supported by laboratory markers including rheumatoid factor, anti-citrullinated protein antibodies, and elevated inflammatory indices. Over recent decades, treatment goals have evolved from symptom control to the prevention of disability and joint damage, and more recently, to achieving early and sustained remission (Smolen et al., 2010; Ajeganova & Huizinga, 2017).

While disease-modifying antirheumatic drugs (DMARDs), including both conventional (tDMARDs) and biologic (bDMARDs) therapies, are effective in controlling disease activity, they are associated with considerable side effects (Maneiro et al., 2014; Verhoef et al., 2017). Emerging evidence shows that a subset of patients can achieve deep remission, immunological remission, and even drug-free remission, suggesting that some patients may effectively be "cure" (Hou et al., 2025; van der Togt et al., 2025). This has motivated strategies for dose reduction and treatment discontinuation, aiming to minimize drug exposure while maintaining disease control (Fautrel et al., 2016; Añez et al., 2024; Nam, 2025).

The effects of RA treatment are typically analyzed using survival methods, which model time-to-event data such as time to flare. Standard survival models assume all patients are susceptible to the event, but when a cured subgroup exists, these models can produce biased estimates (Peng & Yu, 2021). Mixture cure models address this limitation by separating the population into susceptible and cured individuals, allowing the estimation of covariate effects on both the probability of being cured and the timing of events among susceptible patients (Boag, 1949; Ghitany et al., 1994). Although widely applied in oncology (Jia et al., 2013; Felizzi et al., 2021), the application of cure models to RA has been limited.

In this study, we introduce the application of mixture cure models to RA using the R package MDCcure (Monroy-Castillo, Jácome, Cao, Van Keilegom, & Müller, 2025). We analyze clinical trial data in which patients were divided into control and experimental groups, investigating covariates associated with long-term remission and flare risk. This framework enables a more detailed understanding of which patients are likely to maintain remission and which may require ongoing treatment, supporting personalized therapy decisions and optimizing treatment strategies in RA.

2. METHODS

2.1 Mixture cure model

Let Y be the time until the event happens. It is assumed that individuals are subject to random right censoring, and that the censoring time, C, and the time to occurrence of the event, Y, are conditionally independent given a set of covariates $\boldsymbol{X}$. The conditional distribution function of Y is $F(t|\boldsymbol{x}) = P(Y \leq t|\boldsymbol{X} = \boldsymbol{x})$, and the corresponding survival function is $S(t|\boldsymbol{x}) = 1 - F(t|\boldsymbol{x})$. Under right censoring, only the pair (T, δ) is observed, where $T = \min(Y, C)$, and $\delta = \mathbb{1}(Y \leq C)$ is the uncensoring indicator. Moreover, the conditional distribution functions of C and T are $G(t|\boldsymbol{x})$ and $H(t|\boldsymbol{x})$, respectively. On the other hand, the cure indicator is $\nu = \mathbb{1}(Y = \infty)$, with $\nu = 0$ if the individual is susceptible to the event, and $\nu = 1$ otherwise (it is cured). The probability of not being cured (*incidence*) is $p(\boldsymbol{x}) = P(\nu = 0|\boldsymbol{X} = \boldsymbol{x})$, and the conditional survival function of the uncured group, also called *latency*, is $S_0(t|\boldsymbol{x}) = P(Y > t|\nu = 0, \boldsymbol{X} = \boldsymbol{x})$. Then, the mixture cure model can be written as:

$$S(t|\boldsymbol{x}) \quad = \quad 1 - p(\boldsymbol{x}) + p(\boldsymbol{x})S_0(t|\boldsymbol{x}). \tag{1}$$

An important benefit of this model is that it allows covariates to have different influence on cured and uncured patients. So, one of the most important problems to solve is to test if the covariate X or a set of covariates, $\boldsymbol{X}$, has some effect on the cure rate. Note that it is possible to write $\mathbb{E}\{\nu|\boldsymbol{X} = \boldsymbol{x}\} = 1 - p(\boldsymbol{x})$. Motivated by this, Monroy-Castillo, Jácome, Cao, & Keilegom (2025) proposed a hypothesis test for the cure rate based on the martingale difference correlation (MDC) introduced by Shao & Zhang (2014).

2.2 Covariate hypothesis tests based on the martingale difference correlation

The MDC is a natural extension of distance correlation proposed by Szekely et al. (2007), which is used to measure dependence between two vectors. The notion of the MDC is to measure the departure of $(\boldsymbol{X}, Y)$ from the relationship that

$$\mathbb{E}(Y|\boldsymbol{X}) \quad = \quad \mathbb{E}(Y) \quad \text{almost surely,}$$

that is, the conditional mean of Y given $\boldsymbol{X}$ is independent of $\boldsymbol{X}$. The definition of the martingale difference divergence (MDD) was motivated by the definition of the distance covariance. The distance covariance between random vectors $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$ with finite first moments is defined as the non-negative square root of the quantity $\mathcal{V}^2(\boldsymbol{X}, \boldsymbol{Y})$, given by:

$$\mathcal{V}^2(\boldsymbol{X}, \boldsymbol{Y}) \quad = \quad \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{|t|^{1+p}|s|^{1+q}} dt\, ds,$$

where $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$. Similarly, the distance variance, $\mathcal{V}(\boldsymbol{X})$, is defined as $\mathcal{V}(\boldsymbol{X}, \boldsymbol{X})$. We refer the reader to Szekely et al. (2007) for more details on distance covariance and its properties. Based on this framework, the martingale difference divergence (MDD) is defined as follows.

Definition. For random vectors $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$, the martingale difference divergence of Y given $\boldsymbol{X}$ is the nonnegative number $\text{MDD}(Y|\boldsymbol{X})^2$ defined by

$$\text{MDD}(Y|\boldsymbol{X})^2 = \frac{1}{c_p} \int_{\mathbb{R}^p} \frac{|g_{Y,X}(s) - g_Y g_X(s)|^2}{|s|_p^{1+p}} ds, \tag{2}$$

where $g_{Y,X}(s) = \mathbb{E}(Y \exp\{i\langle s, \boldsymbol{X}\rangle\})$, $g_Y = \mathbb{E}(Y)$ and $g_X(s) = \mathbb{E}(\exp\{i\langle s, \boldsymbol{X}\rangle\})$.

Similarly, the martingale difference correlation is defined as follows.

Definition. For random vectors $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$, the martingale difference correlation of Y given $\boldsymbol{X}$ is the non-negative number given by

$$\text{MDC}(Y|\boldsymbol{X})^2 \quad = \quad \begin{cases} \dfrac{\text{MDD}(Y|\boldsymbol{X})^2}{\sqrt{\text{Var}(Y)^2 \mathcal{V}^2(\boldsymbol{X})}}, & \text{Var}(Y)^2 \mathcal{V}^2(\boldsymbol{X}) > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\text{Var}(Y)$ is the variance of Y and $\mathcal{V}^2(\boldsymbol{X})$ is the distance variance of $\boldsymbol{X}$.

Shao & Zhang (2014) showed that $\mathrm{MDC}(Y|\boldsymbol{X}) = 0$ if and only if $\mathbb{E}(Y|\boldsymbol{X}) = \mathbb{E}(Y)$ almost surely.

In the literature, two sample estimators have been proposed for the martingale difference divergence (MDD), analogous to the two estimators introduced for the distance covariance by Szekely et al. (2007) and Huo & Szekely (2016). These two MDD estimators give rise to corresponding estimators of the martingale difference correlation.

The first estimator, referred to as MDDV_n, is based on double-centered matrices and is known to be biased. It was originally proposed by Shao & Zhang (2014) as $\mathrm{MDD}_n(Y|\boldsymbol{X})$. This estimator leads to the corresponding martingale difference correlation estimator, denoted here as MDCV_n, which was also introduced by Shao & Zhang (2014) as $\mathrm{MDC}_n(Y|\boldsymbol{X})^2$. In this approach, the variance of Y is estimated using the sample variance $\mathrm{var}_n(Y)$, and the quantity $\mathcal{V}^2(\boldsymbol{X})$ is estimated using the distance variance dvar_n introduced by Szekely et al. (2007) in Equation (2.9).

An alternative estimator of MDD, denoted as MDDU_n, is a bias-corrected version proposed by Park et al. (2015) as $\widetilde{\mathrm{MDD}}_n(Y|\boldsymbol{X})^2$. It is based on the $\mathcal{U}$-centered matrices defined in Equation (3.5) of the same paper. This estimator leads to a corresponding martingale difference correlation estimator, which we denote by MDCU_n, and is defined as $\widetilde{\mathrm{MDC}}_n(Y|\boldsymbol{X})^2$. Moreover, it has been shown that this estimator is a U-statistic.

These estimators are implemented in the `MDCcure` package through the functions `MDD()` and `MDC()`, where the `center` argument specifies the type of centering applied. Setting `center = "D"` uses double-centering of the matrices, which yields a biased estimator, while `center = "U"` applies U-centering and produces an unbiased estimator.

Both estimators of the MDC, the biased and the bias-corrected versions, are sample-based functions and can therefore also be considered statistics. Based on the proposed statistics, one can construct a MDC test to determine whether the covariate $\boldsymbol{X}$ has an effect on the conditional mean of Y. To approximate the null distribution of the test statistic, two approaches are considered: a *permutation test* and a *chi-square approximation*.

The *permutation test* works by computing the test statistic on the observed data and comparing it to the distribution of the statistic obtained from multiple datasets where the relationship between $\boldsymbol{X}$ and Y has been artificially broken by randomly permuting the values of $\boldsymbol{X}$. This nonparametric method is widely used and described in detail by Gretton et al. (2005) and Pfister et al. (2018).

As a computationally faster alternative, a *chi-square test* is also implemented for the MDCU_n^2 statistic. This approach builds on the asymptotic distribution theory developed by Shen et al. (2022), extending it to the current context. The chi-square approximation offers a more efficient option when computational resources or time are limited. Both ways are implemented in the `mdc_test()` function.

In the same way, Park et al. (2015) proposed a scalar-valued measure to assess the conditional mean independence of Y given $\mathbf{X}$ while controlling for an additional random vector $\mathbf{Z}$, termed the *partial martingale difference correlation* (pMDC). The pMDC provides a natural extension of the partial distance correlation introduced by Szekely et al. (2007), and characterizes the dependency between Y and $\mathbf{X}$ after removing the effect due to $\mathbf{Z}$.

Analogous to how the martingale difference divergence, $\mathrm{MDD}(Y|\mathbf{X})$, is required to define the martingale difference correlation, $\mathrm{MDC}(Y|\mathbf{X})$, the partial martingale difference correlation, $\mathrm{pMDC}(Y|\mathbf{X};\mathbf{Z})$, is defined in terms of the partial martingale difference divergence, $\mathrm{pMDD}(Y|\mathbf{X};\mathbf{Z})$.

Definition. (Remark 4.1 in Park et al. (2015)) Let $Y \in \mathbb{R}$, $\boldsymbol{X} \in \mathbb{R}^p$, and $\boldsymbol{Z} \in \mathbb{R}^r$, and consider $\boldsymbol{W} = (\boldsymbol{X}^T, \boldsymbol{Z}^T)^T \in \mathbb{R}^{p+r}$. The partial martingale difference divergence of Y given $\boldsymbol{X}$, after controlling for the effect of $\boldsymbol{Z}$, is defined as

$$\mathrm{pMDD}(Y|\boldsymbol{X};\boldsymbol{Z}) = \mathrm{MDD}(Y|\boldsymbol{W})^2 - \mathrm{MDD}(Y|\boldsymbol{Z})^2.$$

Before providing the expression of the partial martingale difference correlation (pMDC), we introduce some notation. Consider $C_{\boldsymbol{Z}} = C_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{Z}')$ the (random) double centered version of $c(\boldsymbol{z}, \boldsymbol{z}') = |\boldsymbol{z} - \boldsymbol{z}'|_r$ with respect to $\boldsymbol{Z}$, where

$$\begin{aligned} C_{\boldsymbol{Z}}(\boldsymbol{Z}, \boldsymbol{Z}') &= c(\boldsymbol{Z}, \boldsymbol{Z}') - \int c(\boldsymbol{Z}, \boldsymbol{z}')dF_{\boldsymbol{Z}}(\boldsymbol{z}') - \int c(\boldsymbol{z}, \boldsymbol{Z}')dF_{\boldsymbol{Z}}(\boldsymbol{z}) \\ &\quad + \iint c(\boldsymbol{z}, \boldsymbol{z}')dF_{\boldsymbol{Z}}(\boldsymbol{z})dF_{\boldsymbol{Z}}(\boldsymbol{z}'). \end{aligned}$$

Denote $D_{\boldsymbol{W}}$ the (random) double centered version of $d(\boldsymbol{w}, \boldsymbol{w}') = |\boldsymbol{w} - \boldsymbol{w}'|_{p+r}$ with respect to $\boldsymbol{W}$, and consider

$$R^2(\boldsymbol{Z}, \boldsymbol{W}) = \frac{E(C_{\boldsymbol{Z}} D_{\boldsymbol{W}})}{\sqrt{\mathcal{V}^2(\boldsymbol{Z})\mathcal{V}^2(\boldsymbol{W})}}. \tag{3}$$

Definition. (Proposition 4.1 in Park et al. (2015)) Let $Y \in \mathbb{R}$, $\boldsymbol{X} \in \mathbb{R}^p$, and $\boldsymbol{Z} \in \mathbb{R}^r$, and consider $\boldsymbol{W} = (\boldsymbol{X}^T, \boldsymbol{Z}^T)^T \in \mathbb{R}^{p+r}$. The partial martingale difference correlation (pMDC) of Y given $\boldsymbol{X}$, after controlling for the effect of $\boldsymbol{Z}$ is defined as

$$\text{pMDC}(Y|\boldsymbol{X},\boldsymbol{Z}) = \begin{cases} \frac{\text{MDC}(Y|\boldsymbol{W})^2 - \text{MDC}(Y|\boldsymbol{Z})^2 R^2(\boldsymbol{Z},\boldsymbol{W})}{\sqrt{1-\text{MDC}(Y|\boldsymbol{Z})^4}}, & \text{MDC}(Y|\boldsymbol{Z}) \neq 1, \\ 0, & \text{MDC}(Y|\boldsymbol{Z}) = 1 \end{cases} \tag{4}$$

For a comprehensive treatment of the partial martingale difference correlation and the partial distance correlation, we refer the reader to Park et al. (2015) and Székely & Rizzo (2014), respectively.

2.3 Hypothesis testing for the cure rate in a mixture cure model

Monroy-Castillo, Jácome, Cao, & Keilegom (2025) proposed to test the following hypotheses:

$$\mathcal{H}_0 : \mathbb{E}(\nu|\boldsymbol{X}) \equiv 1-p \quad \text{a.s.} \quad \text{vs} \quad \mathcal{H}_1 : \mathbb{E}(\nu|\boldsymbol{X}) \not\equiv 1-p \quad \text{a.s.}, \tag{5}$$

which tests whether the covariate vector $\boldsymbol{X}$ has an effect on the cure probability. The main problem with the hypotheses (5) is that the response variable (the cure indicator, ν) is only partially observed due to censoring. This challenge was addressed by estimating the cure indicator using the methodology proposed by Amico et al. (2021). The idea was as follows, define τ the upper limit of the support of the lifetime for a susceptible individual, where $\tau = \sup_x \tau(x)$ and $\tau(x) = \inf\{t : S_0(t|x) = 0\}$. It was assumed that $\tau < \infty$ and that the follow-up time is long enough so $\tau < \tau_{G(x)}$ $\forall x$, where $\tau_{G(x)} = \inf\{t : G(t|x) = 1\}$. Therefore, it is reasonable to consider that all the individuals with a censored observed time greater than τ can be categorized as cured ($\nu = 1$).

Since an uncensored subject experiences the event, he/she belongs to the non-cured population with certainty, i.e., $\nu = 0$. Censored observations can be separated into two groups based on the cure threshold τ, and the cure status estimated in a different way for each of them. Those with observed time larger than τ are considered cured, i.e. $\nu = 1$. For the remaining censored observations, the cure status is replaced with the probability:

$$P(\nu = 1|\boldsymbol{X}, C, Y > C) \quad = \quad \frac{1-p(\boldsymbol{X})}{1-p(\boldsymbol{X})+p(\boldsymbol{X})S_0(C|\boldsymbol{X})}.$$

Note that under $\mathcal{H}_0$ in Eq. (5), the cure probability does not depend on $\boldsymbol{X}$, the cure status can be estimated as:

$$\hat{\nu}_h \quad = \quad \mathbb{1}(T > \hat{\tau}) + (1-\delta)\mathbb{1}(T \leq \hat{\tau}) \frac{1-\hat{p}}{1-\hat{p}+\hat{p}\hat{S}_{0,h}(T|\boldsymbol{X})}, \tag{6}$$

where the estimators $\hat{p}$, $\hat{\tau}$ and $\hat{S}_{0,h}(T|\boldsymbol{X})$ are computed as follows. $\hat{\tau}$ is the largest observed uncensored survival time (Xu & Peng, 2014), $\hat{p}$ is the nonparametric estimator proposed by Laska & Meisner (1992) and $\hat{S}_{0,h}(T|\boldsymbol{X})$ is the nonparametric estimator based on the Beran estimator proposed in López-Cheda et al. (2017).

Two statistics for testing covariate effects, based on the MDC between the estimated cure rate $\hat{\nu}_h$ and the covariate $\boldsymbol{X}$, were proposed: $\text{MDCV}_n(\hat{\nu}_h|\boldsymbol{X})^2$ and $\text{MDCU}_n(\hat{\nu}_h|\boldsymbol{X})^2$. Here, MDCV_n and MDCU_n denote the biased and bias-corrected estimators of the MDC, respectively. The null distribution of these statistics was approximated using two approaches, a permutation procedure and a chi-square approximation, yielding three tests: `MDCV`, `MDCU`, and `FMDCU`.

In addition, Monroy-Castillo, Jácome, Cao, & Keilegom (2025) introduced an alternative test based on the statistic

$$\widehat{\mathcal{T}}_n = nh^{1/2} \frac{1}{n} \sum_{i=1}^{n} \{\hat{p}_h(X_i) - \hat{p}\}^2, \tag{7}$$

where $\hat{p}_h(X)$ is the nonparametric estimator of Xu & Peng (2014), and $\hat{p}$ is obtained independently of X using the cure rate estimator of Laska & Meisner (1992). This approach follows the goodness-of-fit framework of Müller & Van Keilegom (2019). The critical values are approximated via the bootstrap procedure described in Section 3 of Müller & Van Keilegom (2019), leading to a fourth test denoted as `GOFT`.

A natural extension arises when attempting to determine whether the cure rate is influenced by specific covariates, when it is known to depend on at least one. To address this issue, Monroy-Castillo, Jácome, Cao, & Keilegom (2025) extended the previous proposal in order to test if the cure rate depends on the covariate $\boldsymbol{X}$ given it depends on the covariate $\boldsymbol{Z}$, as follows:

$$\mathcal{H}_0 : \mathbb{E}(\nu|\boldsymbol{X},\boldsymbol{Z}) \equiv 1-p(\boldsymbol{Z}) \text{ a.s.} \quad \text{vs} \quad \mathcal{H}_1 : \mathbb{E}(\nu|\boldsymbol{X},\boldsymbol{Z}) \not\equiv 1-p(\boldsymbol{Z}) \text{ a.s.} \tag{8}$$

Using the methodology described earlier, the estimator for the cure status in Equation (6) can be extended as follows:

$$\hat{\nu}_{h,\boldsymbol{H}} \quad = \quad \mathbb{1}(T > \hat{\tau}) + (1-\delta)\mathbb{1}(T \leq \hat{\tau}) \, \frac{1 - \hat{p}_h(\boldsymbol{Z})}{1 - \hat{p}_h(\boldsymbol{Z}) + \hat{p}_h(\boldsymbol{Z})\hat{S}_{0,\boldsymbol{H}}(T|\boldsymbol{X}, \boldsymbol{Z})}. \tag{9}$$

For the incidence $p(\boldsymbol{Z})$ consider the estimator $\hat{p}_h(\boldsymbol{Z})$ in Xu & Peng (2014), and for the latency the estimator $\hat{S}_{0,\boldsymbol{H}}(t|\boldsymbol{X}, \boldsymbol{Z})$ which is the extended version of the nonparametric estimator in López-Cheda et al. (2017) to handle two covariates. In particular, a normal multivariate kernel $K_{\boldsymbol{H}}(\boldsymbol{x}) = |\boldsymbol{H}|^{-1/2} K(\boldsymbol{H}^{-1/2}\boldsymbol{x})$ was used with $K(\boldsymbol{x}) = (2\pi)^{-1} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)$, the standard normal density. Here, $\boldsymbol{H}$ denotes the bandwidth matrix, and $|\boldsymbol{H}| = \det(\boldsymbol{H})$.
Based on the partial martingale difference correlation (pMDC) (Park et al., 2015) between the estimated cure status $\hat{\nu}_{h,\boldsymbol{H}}$ and the covariates $\boldsymbol{X}$ conditional on $\boldsymbol{Z}$, a statistic for testing the covariate hypothesis was proposed:

$$\mathrm{pMDC}_n(\hat{\nu}_{h,\boldsymbol{H}}|\boldsymbol{X}, \boldsymbol{Z})^2.$$

The null distribution is approximated using a permutation test. This methodology is implemented in the `testcov2()` function.

2. PATIENTS

A total of 195 patients were enrolled and randomized into the control ($n = 98$) or optimization ($n = 97$) groups. During the first year, 15 flares (15.31%) were observed in the control group compared to 24 flares (24.74%) in the optimization group. In the second year, 140 patients were included in this study. Of these, 73 patients in the control group experienced 12 flares (16.44%) and 67 patients in the optimization group experienced 20 flares (29.85%), see Figure 1.

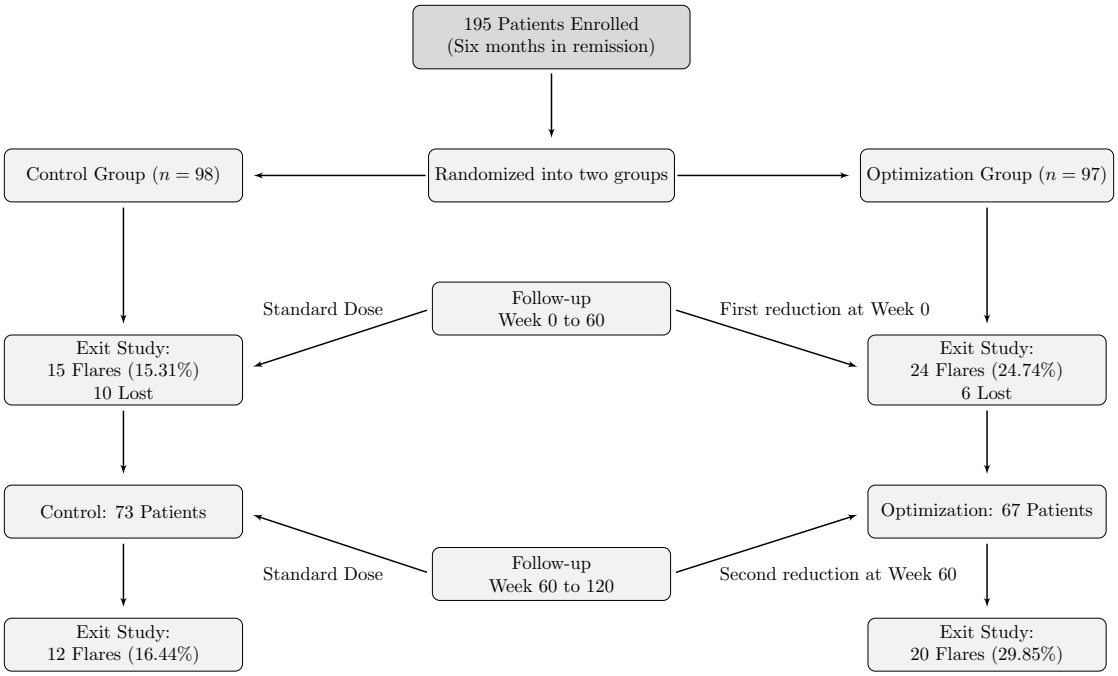

Figure 1: Optimization treatment study diagram.

The clinical database used in this study comprises a rich and diverse collection of variables gathered from a longitudinal cohort of patients enrolled in a randomized clinical trial. The dataset includes administrative and identifying information (e.g., study arm, patient ID), along with detailed sociodemographic characteristics such as age and sex. Clinical and temporal variables are extensively recorded, including the occurrence and timing of disease flares, follow-up visits, and time-to-event data. Anthropometric measurements such as weight, height, and body mass index (BMI) are documented at baseline and throughout follow-up. The dataset also contains vital signs and physiological measurements, including blood pressure, heart rate, respiratory rate, and body temperature. A comprehensive panel of laboratory and biochemical

variables is available, covering hematological, renal, hepatic, lipid, and inflammatory markers, measured both at baseline and during disease activity.

Importantly, the dataset is highly specific to patients with rheumatoid arthritis (RA), as it includes validated disease-specific instruments such as the Health Assessment Questionnaire (HAQ), which assesses functional disability, and widely used clinical disease activity scores such as disease activity score (DAS28) and simplified disease activity index (SDAI), based on physician and patient evaluations, joint counts, and inflammatory markers. These variables provide critical insight into the burden of disease and response to therapy in RA patients. In addition, the database captures genetic information, reproductive health indicators, adverse event reports, and findings from systematic physical examinations. The structure of the dataset allows for longitudinal analysis, with repeated measurements collected at multiple time points (e.g., week 0 to week 120), supporting comprehensive modeling of disease progression, treatment response, and outcome prediction in the context of rheumatoid arthritis.

3. RESULTS

Figure 2 shows the curves that estimate the probability of remaining relapse-free at each time point. In the first year, the survival trajectories of the two groups were similar, suggesting comparable short-term outcomes. By the second year, the curves tended to stabilize at relatively high values, indicating that a substantial fraction of patients in both groups maintained long-term remission. A plateau in the survival curves suggests that a subgroup of patients never experienced the event (relapse/flare), implying the existence of a "cure" fraction.

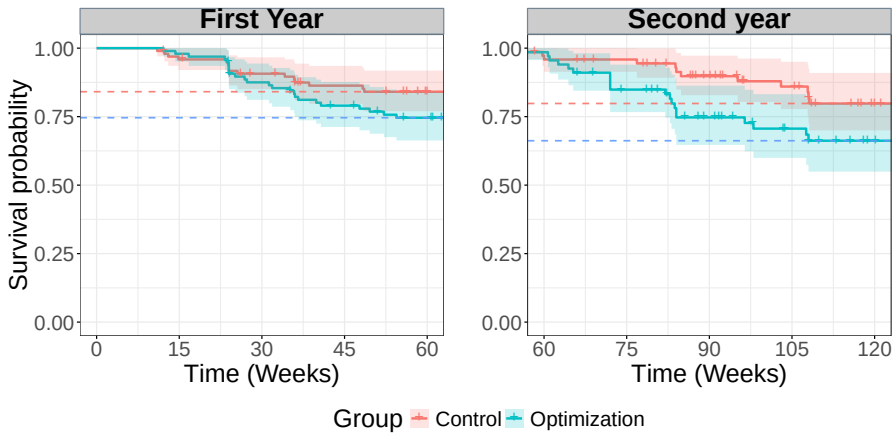

Figure 2: Survival function for each group, control and optimization, compared with their respective cure probability (dashed line).

During the first year, the estimated cure probability was 0.831 for the control group and 0.746 for the optimization group. This indicates that patients receiving standard treatment at baseline had an 83% probability of remaining in remission throughout the first year, whereas those undergoing treatment reduction had a 75% probability.

By the second year, these probabilities decreased further to approximately 80% and 66% for the control and optimization groups, respectively (see Figure 2, Table 1). To assess whether these differences were statistically significant, we compared cure probabilities using the `testcov()` function from the `MDCcure` package. The function returns the four proposals in Monroy-Castillo, Jácome, Cao, & Keilegom (2025), `MDCU`, `MDCV`, `FMDCU` and `GOFT`. Additionally, the function returns the $\hat{\nu}_h$ given in Equation (6).

Although treatment optimization was linked to a lower probability of remission, the difference was not statistically significant (Table 1). In other words, the group covariate showed no significant effect on cure probability. However, by the second year, the gap in cure probabilities widened and the corresponding p-value decreased compared to the first year, suggesting that treatment optimization might be associated with a reduction in cure probability over time. From a clinical perspective, this implies that reducing treatment through an optimization strategy does not meaningfully compromise the chances of achieving or sustaining long-term remission. Nonetheless, it remains essential to identify the covariates that shape this probability within each group.

For example, Figure 3 shows the estimated cure probability as a function of the Disease Activity Score (DAS28) for both groups during the first and second years. In the first year, both groups maintained

	First year			Second year		
	Cure prob.		***p*-value**	**Cure prob.**		***p*-value**
	Cont.	Opt.		Cont.	Opt.	
MDCU	0.841	0.746	0.1689	0.798	0.662	0.1075
MDCV			0.1704			0.1025
FMDCU			0.1631			0.1030
GOFT			0.1491			0.0946

Table 1: Estimated probability of remaining in remission in the control (Cont.) and optimization (Opt.) groups obtained using the unconditional cure rate estimator (Laska & Meisner, 1992) and p-values for each period.

relatively high cure probabilities across the observed DAS28 range, with lower DAS28 values corresponding to a higher likelihood of remaining in remission. In the control group, a DAS28 of 0.485 was associated with a cure probability of 0.967 (95% CI: 0.911–1.00), while a DAS28 of 2.45 reduced the probability to 0.794 (95% CI: 0.698–0.89). Similarly, in the optimization group, a DAS28 of 0.587 corresponded to a cure probability of 0.845 (95% CI: 0.74–0.95), whereas a DAS28 of 2.46 lowered it to 0.708 (95% CI: 0.602–0.815).

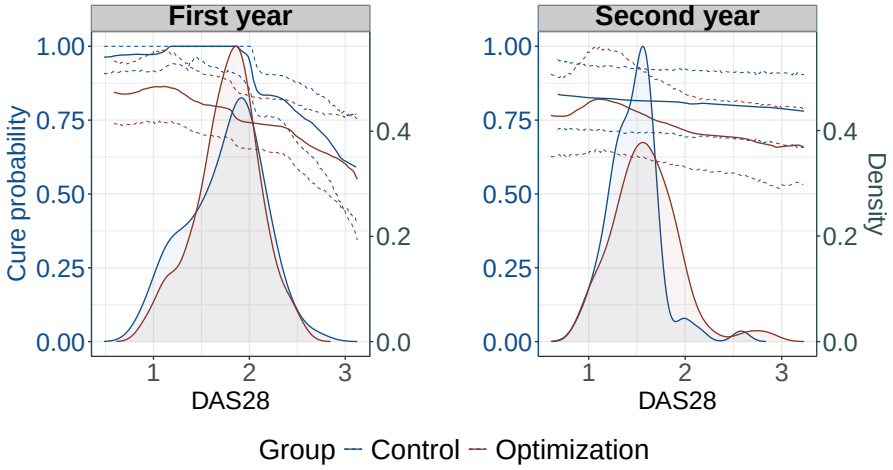

Figure 3: Estimated cure probability and 95% confidence bands conditional on DAS28 and density for the two groups, control (blue) and experimental (red), during first year (left) and second year (right).

Interestingly, the cure probability in the control group appeared to level off over time, indicating that baseline DAS28 strongly influenced early remission but its impact lessened by the second year. In contrast, for the optimization group, baseline DAS28 (week 60) continued to affect the likelihood of remission, as shown by the persistent decrease in remission probability with higher DAS28 values (see Figure 3). These results suggest that DAS28 may play a significant role in predicting sustained remission across both groups over the two-year follow-up.

Nonparametric tests revealed that DAS28, Erythrocyte Sedimentation Rate (ESR), and C-Reactive Protein (CRP) significantly affected the probability of remaining in remission in the control group during the first year, with higher values associated with a notably lower chance of remission (see Table 2).

Whereas in the optimization group, the significant covariates were DAS28-CRP and 3V-DAS28-CRP.

The analysis was initially performed using a univariate approach for each covariate using the `testcov()` function. To account for multiple testing, the False Discovery Rate (FDR) (Benjamini & Hochberg, 1995) correction was subsequently applied, ensuring that the reported significant effects reflect robust associations rather than chance findings.

It is important to emphasize that these tests are nonparametric, meaning that the distribution of the cure probability is not specified and may present certain limitations. To address this, the `MDCcure` package provides tools to assess and visualize whether the estimated cure probability follows common link functions typically used in mixture cure models. In particular, the function `goft()` implements the goodness-of-fit

Control group				
Covariate	MDCU	MDCV	FMDCU	GOFT
CRP	0.0117	0.0117	0.0077	0.0700
ESR	0.0117	0.0117	0.0087	0.1712
DAS28	0.0117	0.0117	0.0077	0.0000
Optimization group				
DAS28-CRP	0.0175	0.0350	0.0141	0.0000
3V-DAS28-CRP	0.0233	0.0350	0.0141	0.0000

Table 2: Covariates having a significant effect on the probability of remaining in remission during the first year, after controlling the false discovery rate (p-values).

test proposed by Müller & Van Keilegom (2019) for three possible link functions: probit, logit, and cloglog. In addition, the function `plotCure()` allows for the visualization of both the semiparametric and nonparametric estimates of the cure probability.

For illustrative purposes, a semiparametric mixture cure model with a logistic link for the cure probability was fitted during the first year, considering DAS28, ESR, and CRP in the control group, and DAS28-CRP and 3V-DAS28-CRP in the optimization group.

In the control group, the data were well described by the logistic model for CRP ($p = 0.880$), ESR ($p = 0.750$), and DAS28 ($p = 0.990$). The semiparametric model (Table 3) further indicated that CRP ($p = 0.0237$) and DAS28 ($p = 0.0317$) had a significant effect on the cure probability, while the effect of ESR was borderline significant ($p = 0.0565$). In the optimization group, the goodness-of-fit test advise against applying a semiparametric cure model with a logistic fit for DAS28-CRP ($p = 0.018$) and 3V-DAS28-CRP ($p = 0.025$). Consequently, we cannot be confident that the logistic model (Table 3) accurately captures the underlying cure dynamics, implying that alternative estimation approaches should be explored. These alternatives might include models based on different distributional assumptions, potentially offering a more precise representation of cure probability.

	Estimate	Std.Error	Z value	$\Pr(>\|Z\|)$
Control group				
	CRP			
θ_0	-2.3661	0.3351	-7.0607	1.66e-12
θ_1	2.1804	0.9642	2.2614	0.0237
	ESR			
θ_0	-2.5745	0.4846	-5.3126	9.16e-08
θ_1	0.0486	0.0254	1.9159	0.0565
	DAS28			
θ_0	-5.3596	1.9436	-2.7576	0.0058
θ_0	1.6255	0.7566	2.1485	0.0317
Optimization group				
	DAS28-CRP			
θ_0	-4.8989	1.1556	-4.2393	2.242e-05
θ_1	2.1963	0.6247	3.5159	4.382e-04
	3V-DAS28-CRP			
θ_0	-5.5511	1.4046	-3.9521	7.748e-05
θ_1	2.5816	0.7614	3.3906	6.975e-04

Table 3: Estimated parameters of the logistic fit of the cure rate for covariates identified as significant by the nonparametric tests, in the control and optimization groups during the first year.

The `plotCure()` function produces Figure 4, which displays the nonparametric and logistic estimates of the cure probability for the covariates DAS28, ESR, and CRP in the control group during the first year. The similarity of the curves across both methods supports the adequacy of the semiparametric mixture cure model. For the logistic estimation, the parameters were obtained using the `smcure()` function from

the `smcure` package (Cai et al., 2012) (see Table 3).

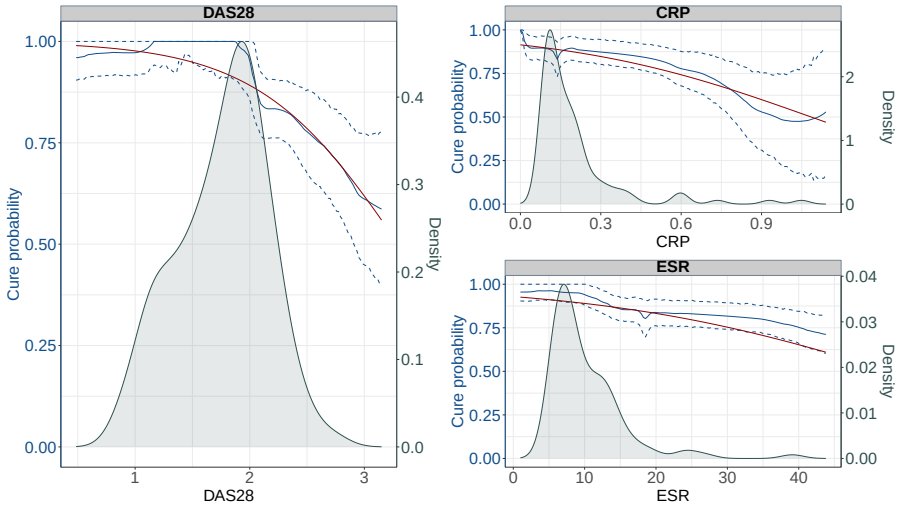

Figure 4: Comparison of nonparametric (blue solid lines) and parametric (red solid line) estimates of cure probability for significant covariates in the control group during the first year of follow-up. The blue dashed lines denote the 95% confidence intervals of the nonparametric estimates. The shaded areas correspond to the empirical density of each covariate.

On the other hand, Figure 5 presents the comparison between the nonparametric and semiparametric estimates of the cure probability for DAS28-CRP and 3V-DAS28-CRP in the optimization group. In contrast to the results obtained for the control group, the discrepancies between the curves are more pronounced, particularly at higher values of the covariates. This suggests that the logistic specification may not adequately capture the functional form of the cure probability in this context. Such deviations reinforce the importance of using flexible nonparametric approaches, as they do not impose restrictive assumptions on the distributional form of the cure probability. From a clinical perspective, these findings indicate that the effect of composite disease activity scores on long-term remission may be more complex and not easily explained by a simple logistic model.

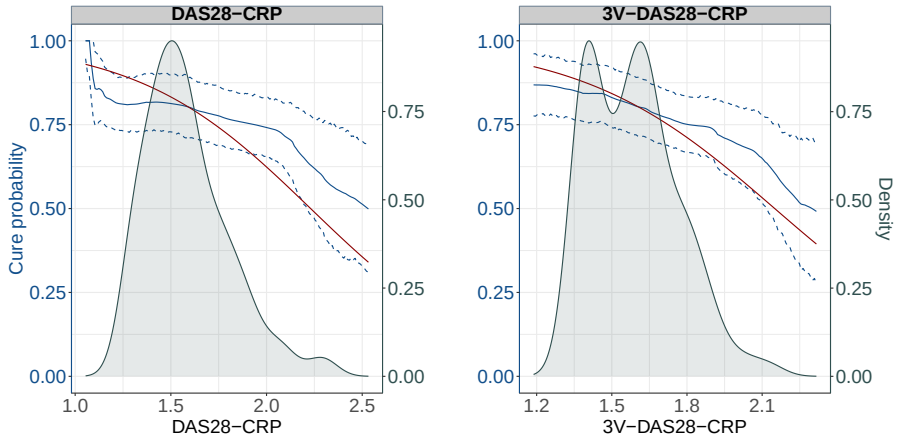

Figure 5: Nonparametric estimation (blue line) of the cure probability compared with the logistic fit (red line).

CONCLUSIONS

This work highlights the usefulness of the `MDCcure` package in R for the analysis of mixture cure models. The package incorporates distance-based measures, such as the martingale difference correlation and the partial martingale difference correlation, providing flexible tools for assessing covariate effects. In

addition, it offers functions for testing whether specific covariates influence the cure probability through four nonparametric hypothesis tests. Complementing this, the package also includes procedures to evaluate the adequacy of different link functions, logit, probit, and cloglog, and a function to compare nonparametric and semiparametric estimates of the cure probability, thereby facilitating both inference and model validation in practical applications.

From the application to the clinical dataset, several relevant findings emerge. First, although treatment optimization was associated with a slightly lower probability of long-term remission compared to standard treatment, this difference did not reach statistical significance. Importantly, the cure probabilities remained relatively high in both groups, suggesting that treatment reduction strategies may be feasible without substantially compromising clinical outcomes. Second, covariates such as DAS28, ESR, and CRP in the control group, and DAS28-CRP and 3V-DAS28-CRP in the optimization group, were identified as significant predictors of remission, underscoring their clinical relevance for individualized treatment assessment during the first year. Third, while semiparametric logistic models provided a good fit for some covariates, marked deviations were observed for others, particularly in the optimization group, highlighting the importance of flexible nonparametric approaches that avoid restrictive assumptions on the functional form of the cure probability.

ACKNOWLEDGMENTS

This research was supported by the International, Interdisciplinary and Intersectoral Information and Communications Technology PhD programme (3-i ICT) granted to CITIC and supported by the European Union through the Horizon 2020 research and innovation programme under a Marie Sklodowska-Curie agreement (H2020-MSCA-COFUND).

REFERENCES

Ajeganova, S., & Huizinga, T. (2017). Sustained remission in rheumatoid arthritis: latest evidence and clinical considerations. *Therapeutic Advances in Musculoskeletal Disease*, *9*(10), 249–262. doi: 10.1177/1759720X17720366

Amico, M., Van Keilegom, I., & Han, B. (2021). Assessing cure status prediction from survival data using receiver operating characteristic curves. *Biometrika*, *108*(3), 727–740. doi: 10.1093/biomet/asaa080

Añez, G., Torrente-Segarra, V., Bonet, M., Vilella, M. C., Orpinell, L., Fernández, A. P., ... De Agustin, J. J. (2024). Clinical and ultrasound optimization in rheumatoid arthritis for patients in sustained remission, can it work as a new optimization tool? *Journal of Ultrasound*, 1–7. doi: 10.1007/s40477-024-00963-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, *57*(1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Boag, J. W. (1949). Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society: Series B (Methodological)*, *11*(1), 15–44. doi: 10.1111/j.2517-6161.1949.tb00020.x

Cai, C., Zou, Y., Peng, Y., & Zhang, J. (2012). smcure: An R-Package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, *108*(3), 1255–1260. doi: 10.1016/j.cmpb.2012.08.013

Fautrel, B., Pham, T., Alfaiate, T., Gandjbakhch, F., Foltz, V., Morel, J., ... others (2016). Step-down strategy of spacing tnf-blocker injections for established rheumatoid arthritis in remission: results of the multicentre non-inferiority randomised open-label controlled trial (strass: Spacing of tnf-blocker injections in rheumatoid arthritis study). *Annals of the Rheumatic Diseases*, *75*(1), 59–67. doi: 10.1136/annrheumdis-2014-206696

Felizzi, F., Paracha, N., Pöhlmann, J., & Ray, J. (2021). Mixture cure models in oncology: a tutorial and practical guidance. *PharmacoEconomics-Open*, *5*, 143–155. doi: 10.1007/s41669-021-00260-z

Ghitany, M., Maller, R. A., & Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis*, *49*(2), 218–241. doi: 10.1006/jmva.1994.1023

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B., & Hyvärinen, A. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research*, *6*(12), 2075–2129.

Hou, Y., Peng, Y., Jin, J., & Li, Z. (2025). Promise of rheumatoid arthritis therapy: From clinical deep remission to drug-free remission. *Best Practice & Research Clinical Rheumatology*, *39*(1), 102031. doi: 10.1016/j.berh.2024.102031

Huo, X., & Szekely, G. J. (2016). Fast Computing for Distance Covariance. *Technometrics*, *58*(4), 435–447. doi: 10.1080/00401706.2015.1054435

Jia, X., Sima, C. S., Brennan, M. F., & Panageas, K. S. (2013). Cure models for the analysis of time-to-event data in cancer studies. *Journal of Surgical Oncology*, *108*(6), 342–347. doi: 10.1002/jso.23411

Laska, E. M., & Meisner, M. J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, *48*(4), 1223–1234. doi: 10.2307/2532714

López-Cheda, A., Cao, R., Jácome, M. A., & Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, *105*, 144–165. doi: 10.1016/j.csda.2016.08.002

Maneiro, J. R., Perez-Pampin, E., Salgado, E., Carmona, L., & Gomez-Reino, J. J. (2014). Observational study of optimization of biologic therapies in rheumatoid arthritis: a single-centre experience. *Rheumatology International*, *34*, 1059–1063. doi: 10.1007/s00296-013-2839-4

Monroy-Castillo, B., Jácome, A., Cao, R., & Keilegom, I. V. (2025). Covariate hypothesis tests for the cure rate in mixture cure models based on martingale difference correlation. *Submitted*.

Monroy-Castillo, B., Jácome, A., Cao, R., Van Keilegom, I., & Müller, U. (2025, jul 23). *MDCcure: Martingale Dependence Tools and Testing for Mixture Cure Models.* https://cran.r-project.org/web/packages/MDCcure/index.html. Retrieved from `https://cran.r-project.org/web/packages/MDCcure/index.html`

Müller, U. U., & Van Keilegom, I. (2019). Goodness-of-fit tests for the cure rate in a mixture cure model. *Biometrika*, *106*(1), 211–227. doi: doi.org/10.1093/biomet/asy058

Myasoedova, E., Crowson, C. S., Giblon, R. E., McCarthy-Fruin, K., Schaffer, D. E., Wright, K., ... Davis, J. M. (2019). Optimization of flare management in patients with rheumatoid arthritis: results of a randomized controlled trial. *Clinical Rheumatology*, *38*, 3025–3032. doi: 10.1007/s10067-019-04664-5

Nam, B. (2025). Optimization of biological therapy selection in rheumatoid arthritis: insights into anemia and interleukin-6 pathway. *Journal of Rheumatic Diseases*, *32*(1), 1–2. doi: 10.4078/jrd.2024.0133

Park, T., Shao, X., & Yao, S. (2015). Partial martingale difference correlation. *Electronic Journal of Statistics*, *9*(1), 1492–1517. doi: 10.1214/15-EJS1047

Peng, Y., & Yu, B. (2021). *Cure models: methods, applications, and implementation.* Chapman and Hall/CRC.

Pfister, N., Bühlmann, P., Schölkopf, B., & Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(1), 5–31. doi: 10.1111/rssb.12235

Shao, X., & Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, *109*(507), 1302–1318. doi: 10.1080/01621459.2014.887012

Shen, C., Panda, S., & Vogelstein, J. T. (2022). The chi-square test of distance correlation. *Journal of Computational and Graphical Statistics*, *31*(1), 254–262. doi: 10.1080/10618600.2021.1938585

Smolen, J. S., Aletaha, D., Bijlsma, J. W., Breedveld, F. C., Boumpas, D., Burmester, G., ... others (2010). Treating rheumatoid arthritis to target: recommendations of an international task force. *Annals of the Rheumatic Diseases*, *69*(4), 631–637. doi: 10.1136/ard.2009.123919

Székely, G. J., & Rizzo, M. L. (2014). Partial Distance Correlation with Methods for Dissimilarities. *The Annals of Statistics*, *42*(6), 2382–2412. doi: 10.1214/14-AOS1255

Szekely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794. doi: 10.1214/009053607000000505

van der Togt, C. J., den Broeder, N., Boonstra, M. S., van der Maas, A., den Broeder, A. A., & van Herwaarden, N. (2025). Disease activity–guided dose optimization including discontinuation of tnf inhibitors in rheumatoid arthritis is effective for up to 10 years: an observational follow-up of the DRESS study. *Rheumatology*, *64*(2), 533–540. doi: 10.1093/rheumatology/keae103

Verhoef, L. M., Tweehuysen, L., Hulscher, M. E., Fautrel, B., & den Broeder, A. A. (2017). bDMARD dose reduction in rheumatoid arthritis: a narrative review with systematic literature search. *Rheumatology and Therapy*, *4*, 1–24. doi: 10.1007/s40744-017-0055-5

Wasserman, A. M. (2011). Diagnosis and management of rheumatoid arthritis. *American Family Physician*, *84*(11), 1245–1252.

Xu, J., & Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, *42*(1), 1–17. doi: 10.1002/cjs.11197