



# Predicting donations and profiling donors in a blood collection center: a Bayesian approach

Ilenia Epifani<sup>1</sup> · Ettore Lanzarone<sup>2</sup> · Alessandra Guglielmi<sup>1</sup> 

Accepted: 23 October 2023 / Published online: 10 November 2023  
© The Author(s) 2023

## Abstract

Donor profiling and donation prediction are two key tasks that any blood collection center must face. Profiling is important to target promotion campaigns, recruiting donors who will guarantee a high production of blood units over time. Predicting the future arrivals of donors allows to size the collection center properly and to provide reliable information on the future production of blood units. Both tasks can be addressed through a statistical prediction model for the intensity function of the donation event. We propose a Bayesian model, which describes this intensity as a function of individual donor's random frailties and their fixed-time and time-dependent covariates. Our model explains donors' behaviors from their first donation based on their individual characteristics. We apply it to data of recurrent donors provided by the Milan department of the *Associazione Volontari Italiani del Sangue* in Italy. Our method proved to fit those data, but it can also be easily applied to other blood collection centers. The method also allows general indications to be drawn, supported by quantitative analyses, to be provided to staff.

**Keywords** Bayesian model · Blood donations · Heterogeneity · Prediction · Recurrent events

---

✉ Alessandra Guglielmi  
alessandra.guglielmi@polimi.it

Ilenia Epifani  
ilena.epifani@polimi.it

Ettore Lanzarone  
ettore.lanzarone@unibg.it

<sup>1</sup> Department of Mathematics, Politecnico di Milano, 20133 Milan, Italy

<sup>2</sup> Department of Management, Information and Production Engineering, University of Bergamo, 24044 Dalmine, BG, Italy

## 1 Introduction

Human blood is a key component for several care treatments and plays a crucial role in all health care systems. It is needed to save lives in acute emergencies, to allow for many types of surgical interventions, such as organ transplants, and it is continuously required for the survival of chronic patients. Unfortunately, blood cannot be produced in laboratory but can only be withdrawn from healthy subjects, and its short shelf life limits the period between withdrawal and use. Therefore, blood is a limited resource, while its demand is very high. For example, before COVID-19, the demand was about 10 million units per year in the US and 2.1 million in Italy (World Health Organization 2012), and these values are growing again. In Western countries, blood is usually collected from volunteer *donors*, i.e., unpaid individuals who donate their blood voluntarily and for free. Blood is classified into groups and according to the Rhesus factor (Rh), and patients receive the blood of their own type (combination of group and Rh factor) or a compatible one. There are two types of donations, whole-blood and apheresis, where apheresis refers to the donation of specific blood constituents, such as erythrocytes or platelets, in which a mechanical apparatus separates the required blood constituents and reinfuses the others into the donor.

Blood is supplied by the Blood Donation (BD) system, which is tasked with providing an adequate supply of blood units to meet the demand of transfusion centers and hospitals, while respecting their storage capacity and the temporal profile of the demand. The BD Supply Chain (BDSC) can be divided into four echelons (Sundaram and Santhanam 2011): collection, transportation, storage and utilization. Many problems arise in the BDSC management from the collection echelon to the final utilization of blood units, which have been largely addressed in the literature (Beliën and Forcé 2012; Baş et al. 2016). In this paper, we focus on the collection echelon, which is very relevant to the entire BDSC because problems at this stage may deteriorate the performance of the entire BDSC and impact blood shortages and wasted units (Baş Güre et al. 2018). For example, it is straightforward that increasing the number of donations improves the performance of the BD system, but also an effective management of the available donations, which directs donors to suitable days, can avoid shortage and wastage of blood units.

The management of a blood collection center must take into account a twofold perspective (Baş Güre et al. 2018; Baş et al. 2018). On the one hand, it should pursue operational goals common to several health care providers, such as waiting time reduction, optimal workforce planning, and effective appointment scheduling. On the other hand, a blood collection center produces blood units and blood products to meet storage and utilization demands. Two aspects are particularly critical from the production viewpoint. Firstly, the number of produced units and all activities carried out at any blood collection center strongly depend on the number of donors who arrive daily at the center to make a donation. Predicting the daily number of donors in advance is therefore essential for proper planning and sizing of the collection center, and to provide reliable information on

the future production of blood units to the rest of the BDSC echelons. Secondly, blood collection centers invest, also in economic terms, to carry out campaigns to promote and acquire further donors. The goal is to enroll novel donors who regularly and frequently donate blood. Therefore, it is important to identify the most productive donor profiles to better target these promotion campaigns and recruit donors who will guarantee a high production of blood units over time. Both needs can be addressed through a statistical prediction model for the intensity function of the donation event. For the *profiling* goal, the profiles of donors who donate more frequently can be identified by analyzing donors' characteristics that significantly yield shorter waiting times before the next donation. For the *prediction* goal, the number of donors arriving on each day of a given horizon can be obtained by combining the predictions of all donors who can donate within that horizon.

This work considers whole-blood donations, which cover most of the donation events. We propose a Bayesian model for the intensity function of the blood donation event, which describes it as a function of individual donor's random frailties and their time-dependent covariates. The aim is to explain donors' behaviors since their first donation, based on their individual characteristics. Under the Bayesian approach, the parameters of the likelihood (the conditional distribution of individuals in the sample) are random, and all the statistical inference is based on the posterior distributions of these random parameters, namely, the conditional distribution of the parameters given observed data. The posterior distribution of all parameters also allows to predict donors' arrivals, thus supporting planning and other management tasks, and to identify the parameters that influence the intensity of the donation, thus supporting profiling. To show an application of our model, we apply it to the data provided by the Milan department of the *Associazione Volontari Italiani del Sangue* (AVIS), referred to as AVIS Milan in the following. AVIS is the largest Italian blood donor organization, founded in Milan in 1927. Today, it ensures about 80% of the national blood supply, and is present throughout the country with over 3400 centers. AVIS Milan, one of the most important nodes in the AVIS network, collects blood from donors residing or working in Milan. It collects about 1500 whole-blood donations per month and supplies them to the Niguarda hospital, located in the city. The dataset includes the list of donations from each donor, along with the measurements of some donors' vital parameters (e.g., heart rate, blood pressure and hemoglobin) and information on donor' habits acquired before each donation through an interview with a physician.

The remainder of this paper is organized as follows. Section 2 overviews the literature related to the problem addressed in this work. Section 3 describes the variables and provides an exploratory analysis of the available dataset. Section 4 details our model, while Sect. 5 shows the posterior inference. Then, Sect. 6 exploits the model outcomes and predictions to support profiling decisions and management tasks. Finally, Sect. 7 concludes the work.

## 2 Literature review

The BDSC has been extensively studied, as documented in Osorio et al. (2015) and Başı Güre et al. (2018). However, these literature reviews show that most studies focus on the storage and utilization echelons of the BDSC. In contrast, the collection has been marginally studied compared to the others (Ayer et al. 2018, 2019; Başı et al. 2018).

The work of Başı et al. (2018) has contributed first to the development of a decision support tool for blood collection, proposing an appointment scheduling system that includes a linear programming model for preallocating time slots to blood types and a prioritization policy to assign a slot when a donor makes the reservation. Then, this framework was extended in Doneda et al. (2023) to include home blood donations and in Yalçındağ et al. (2020) to face uncertain donor arrivals. In fact, uncertainty is recognized as a major sticking point when dealing with blood donation management (Lanzarone and Yalçındağ 2019), as well as in several health care management contexts (Addis et al. 2015). Therefore, it must necessarily be considered to adequately address the problems arising in blood donation management, and to create more effective tools, solution methods and decision support systems for BDSC. In the following, to show its impact, we first overview recent works that deal with uncertainty in the management of blood collection. Then, we analyze recent works that propose stochastic models to predict the uncertain quantities affecting blood donation and to classify donors.

From the management viewpoint, Jabbarzadeh et al. (2014) developed a robust optimization model for blood facility location and allocation decisions during post-disaster periods under supply and demand uncertainty. Zahiri et al. (2015) adopted a robust possibilistic fuzzy programming approach to determine the best locations of blood facilities coping with several uncertain parameters. Ramezani and Behboodi (2017) developed a robust optimization approach for the location-allocation problem of blood collection centers in the presence of stochastic demands. Rabbani et al. (2017) analyzed the mobile blood collection system for platelet production with uncertain donors' arrivals. Hamdan and Diabat (2019) proposed a stochastic model for red blood cell supply that simultaneously considers production, inventory and location decisions. Finally, as mentioned, Yalçındağ et al. (2020) included uncertain donors' arrivals in the BD appointment scheduling.

Other works specifically focused on prediction and classification tasks in BD (Khalid et al. 2013). Darwiche et al. (2010) combined a principal component analysis and a support vector machine to predict blood donation occurrences, and applied this combined approach to donor data from a blood transfusion service in Taiwan. Santhanam and Sundaram (2010) and Sundaram and Santhanam (2011) used decision trees to classify donors, in order to determine voluntary blood donorship based on blood donation patterns. Ramachandran et al. (2011) classified blood donors using a decision tree to identify regular donors and enable blood banks to organize blood donation camps efficiently. Similarly, Boonyasith and Jittamai (2012) used neural networks and decision trees, to identify

patterns in blood donors' behaviors based on the factors influencing the donation decision. In a perspective more similar to that of our work, Testik et al. (2012) adopted a two-step cluster method, together with classification and regression trees, to identify donors' daily and hourly arrival patterns, considering data from a Turkish hospital. Khalilinezhad et al. (2014) used association rule mining to find the best donors within the whole population, and applied their approach to data from two cities in the Middle East. More recently, Alkahtani and Jilani (2019) adopted a classification approach to predict returning donors and time series analysis to predict donation dates, focusing on the lower number of returning donors versus the higher number of non-returning ones. Bischoff et al. (2019) used time series forecasting to predict the daily number of donations to a tertiary care center, to account for a decrease in platelets production preemptively. Shashikala et al. (2019) applied naive Bayes technique and  $K$ -nearest neighbors algorithm to predict whether individuals are donors or not. Kircic et al. (2020) used logistic regression and a naive Bayes classifier to determine blood donation probabilities. Kauten et al. (2021) applied machine learning algorithms to model donor retention to support cost-effective outreach programs. They focused on predicting which donors will donate blood during a future time window, and applied the algorithm to operational data obtained from a large regional blood center in the US.

A few works addressed the BD prediction task in the Bayesian setting. Tavakol et al. (2016) proposed a log-normal hazard model with gamma correlated frailties to model the chance of donating blood. They considered data from an Iranian province and identified the types of donors with higher chances to donate. Mohammadi et al. (2016) implemented a bivariate zero-inflated Poisson regression to jointly model the number of blood donations and that of blood deferrals. They used non-informative priors, both in the presence and absence of covariates. Kassie and Birara (2021) adopted a Bayesian binary logistic regression approach to assess the impact of the covariates in blood donation, focusing on data from Northwestern Ethiopia.

Differently from these works, we consider recurrent events from the blood donation process. In the statistical literature, recurrent event data are tackled alternatively as: (i) modelling the intensity function of the event counts  $\{N(t), t \geq 0\}$ ; (ii) modelling the whole sequence of gap times between successive realizations of the recurrent events (Cook and Lawless 2007). The second approach is more appropriate when the events are relatively infrequent or when, after an event, individual renewal takes place in some way. The first approach is most suitable when individuals frequently experience the event of interest, as in our application, and the occurrence does not alter the process itself. The canonical framework for analyzing event counts is that of inhomogeneous Poisson processes. Among the original contribution of this paper, we model the intensity function of the whole-blood donation event process as a function of the individual donors' random frailties and their time-dependent covariates. We adopt a full Bayesian approach, assuming a prior distribution for all unknown likelihood parameters.

### 3 Data and variables

We consider donation data of AVIS Milan from January 1st, 2010 to June 30th, 2018 concerning the *recurrent donors*, namely the donors who started to donate after January 1st, 2010 and donated at least two times in those years. Only whole-blood donations are included in the study, and time is measured in days. Furthermore, as the focus is on recurrences, the first donations corresponding to time  $t = 0$  are removed.

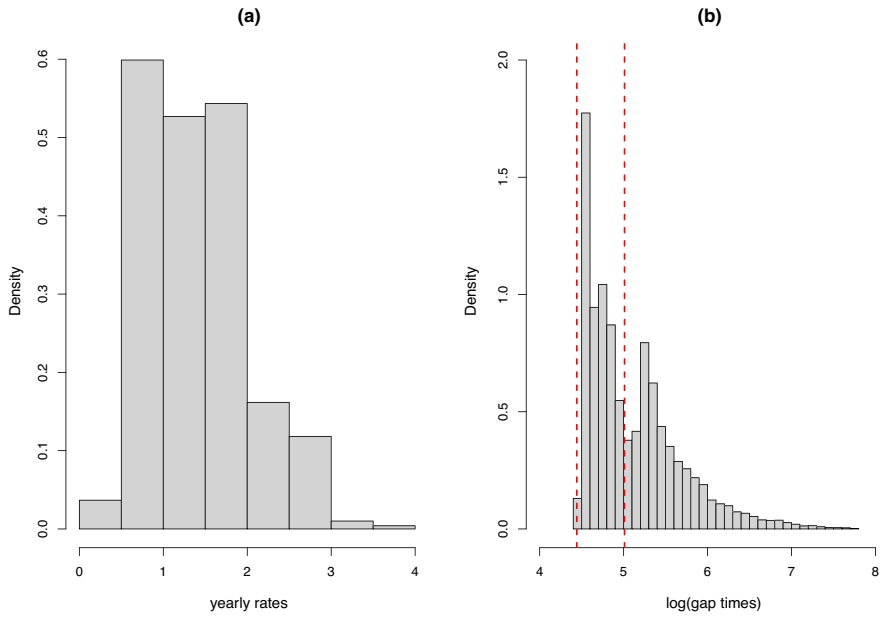
We include two types of data: donations and donors' personal data, and information about donors' habits. Specifically, we include donor ID, sex, age at first donation, age at current donation, blood group (A, B, AB or O) and Rh factor (POS or NEG), donation ID and time of the donation, indicators of smoking, drinking, physical activity and stress level, tea and coffee consumption, diet type, height (in m), weight (in kg), Body Mass Index (BMI), and health state values such as systolic and diastolic blood pressure (SBP and DBP, respectively), heart rate (HR) and hemoglobin (HGB). The health state values are recorded with each donation, measured by a physician. According to the Italian donation rules, any candidate donor who is going to donate for the first time must be between 18 and 60 years old, while the age limit is extended to 65 years for successive donations; however, physicians can allow a donor to keep donating until 70 years old if eligible after clinical evaluation of the age-related risks. Donor's weight must be greater than 50 kg, and blood pressure, heart rate and hemoglobin values must lie between fixed limits. As an example, hemoglobin range is [13, 18] g/dl for male donors and [12, 16] g/dl for female donors. The minimum time gap time between two consecutive donations is 90 days for men and women in menopause, and 180 days for the other women. However, a small tolerance on these thresholds is possible after clinical evaluation. In particular, in our dataset, the minimum gap time is 85 days for men and 150 days for women. These rules are consistent with those that regulate blood donation processes in other countries, such as in Spain (Aldamiz-Echevarria and Aguirre-Garcia 2014). An observation time interval is associated to each donor, starting with the donor's first donation ( $t = 0$ ) and ending on June 30th, 2018. Therefore, the duration of a donor's observation time interval can be different from that of the others. The final dataset consists of 25, 689 observations and 5, 937 unique donors, including 4, 005 men and 1, 932 women.

Donation data show that most recurrent donors made only one donation yearly. Figure 1a displays the histogram of empirical yearly rates of donation, i.e., the total number of donations divided by the number of years under observation for each donor, while Fig. 1b reports the histogram of all donors' gap times. The histogram in Fig. 1a is right-skewed, whereas that in Fig. 1b is bimodal due to the different donation rules between men and women.

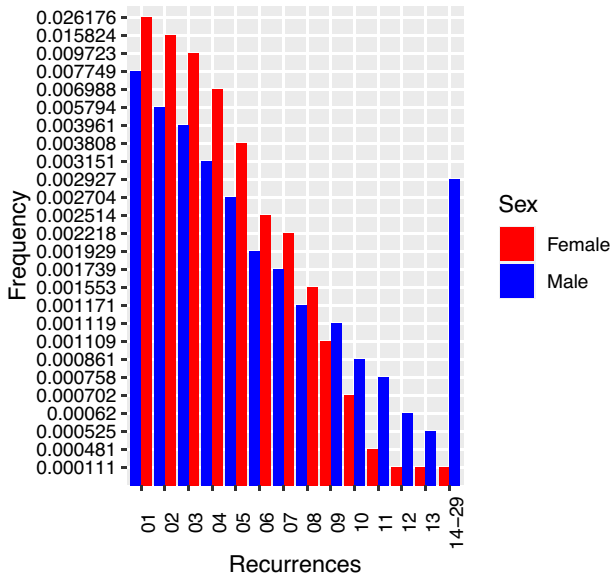
The maximum number of recurrences is 29 for the men and 14 for women. Figure 2 provides the barplots of normalized recurrences grouped by sex, which are defined as:

$$\frac{\text{number of men who perform } k \text{ recurrences}}{\text{total number of men} \times 29}, \quad k = 1, \dots, 29$$

for men and



**Fig. 1** Histograms of yearly rates of donation **(a)** and gap times **(b)** on the log scale. The red vertical dashed lines denote the minimum waiting times for men ( $\log 85$ ) and women ( $\log 150$ ), respectively



**Fig. 2** Barplot of donation normalized recurrences for male and female donors

$$\frac{\text{number of women who perform } k \text{ recurrences}}{\text{total number of women} \times 14}, \quad k = 1, \dots, 14$$

for women. The normalization allows for a fair comparison of female and male donation recurrences. Note that the 99% empirical quantile for women is 11, which coincides with the 88% empirical quantile for men. This means that, subject to law obligations, men donate much less than women.

The available data have missing values in some covariates. Due to a significant number of missing data, diet and stress factors have been discarded from the dataset. Missing values of the other variables were imputed using suitable frequentist methods, via the R package MICE (van Buuren and Groothuis-Oudshoorn 2011). In particular, before imputation, there were 40% missing values for coffee or tea consumption, 14% for HGB, 9% for SBP, DBP and HR, 0.25% for BMI, and 0.19% for smoking, alcohol consumption and physical activity. The peak of missing data was recorded in 2013 and the minimum in 2018.

Among the factors that may affect blood donation recurrence, we investigate the covariates reported in Table 1. Note that information on smoking, alcohol consumption and level of physical activity are communicated to the physician by the donors themselves; as such, they could be inaccurate.

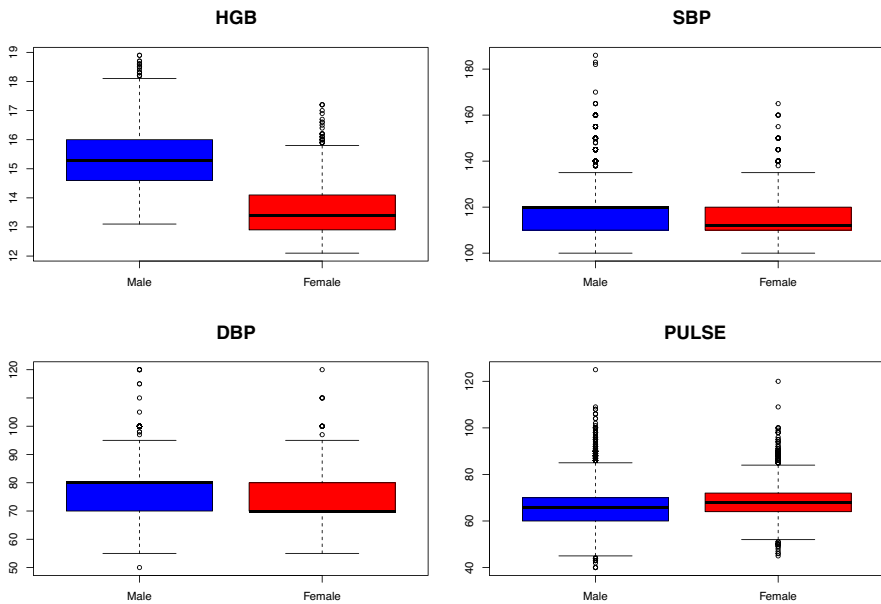
The mean age of the sample at the first donation is 32 years (with a large standard deviation  $\simeq 10$  years), male donors are about twice as female donors, the majority of the population has blood group 0 (46.4%), and positive Rh factor is more frequent than negative Rh factor (about 87 and 84% of male and female donors are

**Table 1** Covariates included in the model

Name	Type	Description
SEX	Binary	Sex: 1 male; 0 female
AGE	Num	Age at the time of first donation
SMOKE	Binary	Smoker: 1 yes; 0 no
ALCOHOL	Binary	Alcohol consumption: 1 yes; 0 no
PHYSICAL_ACTIVITY	Binary	Physical activity: 1 yes; 0 no
RH	Binary	Rh factor: 1 positive; 0 negative
GROUP_0	Binary	Blood group 0
GROUP_A	Binary	Blood group A
GROUP_B	Binary	Blood group B
GROUP_AB	Binary	Blood group AB
BMI	Num	BMI
HR	Num	Heart rate
SBP	Num	Systolic blood pressure
DBP	Num	Diastolic blood pressure
HGB	Num	Hemoglobin
SEX*HGB	Num	Interaction sex - hemoglobin
SEX*RH	Binary	Interaction sex - Rh factor
SEX*BMI	Num	Interaction sex - BMI

**Table 2** Sample frequencies of donors' BMI, classified using the standard weight status categories (Center of Disease Control and Presentation; [www.cdc.gov/healthyweight/assessing/bmi/adult\\_bmi/index.html](http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html))

BMI	Weight status	Tot percentage	Sex	Percentage
< 18.5	Underweight	1.44	F	64.71
			M	35.29
18.5–24.9	Healthy weight	69.00	F	32.29
			M	67.71
25.0–29.9	Overweight	24.84	F	18.49
			M	81.51
≥ 30.0	Obese	4.72	F	27.70
			M	72.30



**Fig. 3** Boxplots of HGB (top left), SBP (top right), DBP (bottom left) and HR (bottom right) grouped by donor sex

Rh-positive, respectively). There are more non-smokers (67.49%) than smokers, alcohol non-consumers (69.82%) than consumers, and 76.03% of the donors practice physical activity. About 69% of donors have healthy weight, underweight women are about twice as underweight men, whereas overweight men are four times overweight women (Table 2).

With regard to the health state variables (SBP, DBP, HR, HGB), the boxplots in Fig. 3 show that women have a lower value of hemoglobin than men, as expected, while SBP, DBP and HR are homogeneous across sexes.

Only HGB and DBP are considered time-dependent variables in the model. This assumption arises from preliminary analyses, which indicated that the trajectories of SBP and HR do not seem to vary over the observation period.

## 4 Methods

Let  $I$  be the number of donors. For each donor  $i = 1, \dots, I$ , we consider a single recurrent event process starting at day  $T_{i0} = 0$  of their first donation, where  $0 < T_{i1} < T_{i2} < \dots$  denote donor  $i$ 's days of subsequent donations. A counting process  $\{N_i(t)\}_{t \geq 0}$  with  $N_i(t) = \sum_{k=1}^{+\infty} \mathbf{1}\{T_{ik} \leq t\}$  records the number of donations of donor  $i$  up to day  $t$  beyond the first donation.  $\{N_i(t)\}_{t \geq 0}$  is right-continuous, and mathematically defined by its event intensity function:

$$\lambda_i(t) = \lim_{\Delta t \rightarrow 0} \frac{P(N_i(t + \Delta t^-) - N_i(t) = 1)}{\Delta t} \quad (1)$$

where the difference  $N_i(t + \Delta t^-) - N_i(t)$  represents the number of donations of donor  $i$  in the interval  $[t, t + \Delta t)$ . Roughly speaking,  $\lambda_i(t)$  gives the instantaneous probability of an event occurring on day  $t$ . Because of (1),  $\{N_i(t)\}_{t \geq 0}$  is an inhomogeneous Poisson process (Cook and Lawless 2007, Chapter 1). Donor  $i$  is observed over the time interval  $[0, c_i]$ , in which he/she donates blood at days  $t_{i1}, \dots, t_{in_i}$  for a total of  $n_i$  donations, with  $0 < T_{i1} = t_{i1} < \dots < T_{in_i} = t_{in_i} \leq c_i$ . If  $t_{in_i} < c_i$  then  $c_i$  is a censoring time.

Donor-specific information is fed into the model by including covariates and an individual random effect function into the associated individual *multiplicative* event intensity function  $\lambda_i(t) = \lambda_i(t|\mathbf{x}_i)$ , modeled as follows:

$$\lambda_i(t|\mathbf{x}_i) = u_i(t) e^{\mathbf{x}'_i(t)\boldsymbol{\beta}} \quad (2)$$

where  $\mathbf{x}'_i(t) = (x_{i1}(t), \dots, x_{ip}(t))$  is the  $p$ -dimensional vector of covariates of donor  $i$  at time  $t$ ,  $\boldsymbol{\beta}$  the vector of regression coefficients, and  $u_i(t)$  the  $i$ -specific donor random effect. Symbol  $'$  denotes the transposition of the column vector. Our model belongs to the class of Cox's proportional hazard regressions with random effects (Klein and Moeschberger 2003). Hence, Eq. (2) can be explained as follows: the logarithm of the intensity  $\lambda_i(t)$  of the  $i$ -th donor is equal to the linear predictor  $\mathbf{x}'_i(t)\boldsymbol{\beta}$  plus an individual random effect  $u_i(t)$ . In this way, the effect of the  $j$ -th covariate is represented by parameters  $\beta_j$ . According to random effects models, the individual parameter  $u_i(t)$  might represent the individual variability in the log-intensity, which is not explained by the covariates.

Some covariates in  $\mathbf{x}_i(t)$  are time-dependent, measured at day  $t$  in conjunction with donations, while others are recorded at day  $t = 0$  and considered constant over time. The time-varying covariates in  $\mathbf{x}_i(t)$  are assumed to be step functions:

$$\mathbf{x}_i(t) = \mathbf{x}_i(t_{ij}) \quad \text{if } t_{ij} \leq t < t_{i,j+1}, \quad j = 0, 1, \dots, n_i, \quad i = 1, \dots, I \quad (3)$$

with  $t_{in_i+1} = c_i$  and  $t_{i0} = 0$  for all  $i$ . The vector  $\mathbf{x}_i(t)$  also includes interaction terms. In this application, the preliminary data analysis (Sect. 3) suggested to include the interactions between sex and hemoglobin, sex and Rh factor, and sex and BMI. The covariates are listed in Table 1. Note that, because of the coding of the blood group, the intercept includes the regression parameter corresponding to `GROUP_A`. Thus, the total number of covariates including interactions is 17. Only HGB and DBP are

considered time-dependent, as mentioned in Sect. 3, and all numerical covariates have been standardized.

To write the likelihood, we first specify the assumptions about the random effect function  $u_i(t)$  of donor  $i$ . We assume a piecewise constant  $u_i(t)$  on the time domain  $[0, c]$ , with  $c = \max_{i=1, \dots, I} c_i$  and  $K = 10$  intervals identified by the cut-points  $0 = a_0 < a_1 < \dots < a_K = c$ , i.e.:

$$u_i(t) = \sum_{k=1}^K u_{ik} \mathbf{1}_{(a_{k-1}, a_k]}(t). \tag{4}$$

As donors cannot donate more than a certain number of times per year, we introduce an at-risk process  $\{Y_i(t)\}_{t \geq 0}$  in the model, which represents the risk of donor  $i$  of experiencing a donation at day  $t$ , with:

$$Y_i(t) = \mathbf{1}_{(T_{N_i(t^-)} + \Phi_i, c_i)}(t). \tag{5}$$

The quantity  $Y_i(t)$  forces the intensity to be 0 for the next  $\Phi_i$  days after every donation, i.e., it imposes that a donor cannot donate until  $\Phi_i$  days after the last donation, and even after their censoring time  $c_i$ .  $\Phi_i$  depends on donor's sex, and we fix  $\Phi_i = 85$  if  $i$  is a male donor and  $\Phi_i = 150$  if  $i$  is a female donor; see the discussion in Sect. 3. Furthermore, as this process only includes the administrative-censored time  $c_i$  and  $\Phi_i$ , it seems reasonable to assume that the at-risk indicator  $Y_i(t)$  and the observation at day  $t$ , given by  $N_i(t + \Delta t^-) - N_i(t)$ , are independent. The contribution  $\mathcal{L}_i$  of donor  $i$  to the likelihood function is derived using Theorem 2.1 in Cook and Lawless (2007), which gives the conditional probability density of  $n_i$  events occurring at days  $t_{i1} < \dots < t_{in_i}$  for each donor  $i$  who recurrently donates over the time interval  $(0, c_i]$ . Hence:

$$\begin{aligned} \mathcal{L}_i &= \prod_{j=1}^{n_i} \lambda_i(t_{ij} | \mathbf{x}_i(t)) \exp \left\{ - \int_0^{c_i} Y_i(s) \lambda_i(s | \mathbf{x}_i(s)) ds \right\} \\ &= \prod_{j=1}^{n_i} u_i(t_{ij}) \exp \left\{ \mathbf{x}'_i(t_{ij}) \boldsymbol{\beta} - \int_0^{c_i} Y_i(s) u_i(s) e^{\mathbf{x}'_i(s) \boldsymbol{\beta}} ds \right\} \\ &= \prod_{k=1}^K u_{ik}^{n_{ik}} \exp \left\{ \sum_{j=1}^{n_i} \mathbf{x}'_i(t_{ij}) \boldsymbol{\beta} - \sum_{k=1}^K u_{ik} \int_{a_{k-1}}^{a_k} Y_i(s) e^{\mathbf{x}'_i(s) \boldsymbol{\beta}} ds \right\} \end{aligned} \tag{6}$$

where  $n_{ik}$  is the number of donations experienced by donor  $i$  in the interval  $(a_{k-1}, a_k]$  and  $n_i = \sum_{k=1}^K n_{ik}$ . Consequently, the likelihood of all  $I$  donors is:

$$\mathcal{L} = \prod_{i=1}^I \mathcal{L}_i = \left( \prod_{i=1}^I \prod_{k=1}^K u_{ik}^{n_{ik}} \right) \exp \left\{ \sum_{i=1}^I \sum_{j=1}^{n_i} \mathbf{x}'_i(t_{ij}) \boldsymbol{\beta} - \sum_{i=1}^I \sum_{k=1}^K u_{ik} \int_{a_{k-1}}^{a_k} Y_i(s) e^{\mathbf{x}'_i(s) \boldsymbol{\beta}} ds \right\}. \tag{7}$$

As for the prior, we assume *a priori* independence between  $\boldsymbol{\beta}$  and  $\{u_{ik}\}$  with:

$$\beta_1, \dots, \beta_p \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2) \tag{8}$$

and a hierarchical gamma prior for the steps  $\{u_{ik}\}$ , with hyperparameters  $\{\lambda_{0k}\}$ , which takes into account the length of the time-interval  $a_k - a_{k-1}$  (Christensen et al. 2010):

$$\begin{cases} u_{ik} \mid \lambda_{0k}, c \stackrel{\text{i.i.d.}}{\sim} \text{Gamma} (c(a_k - a_{k-1})\lambda_{0k}, c(a_k - a_{k-1})) \\ \lambda_{0k} \stackrel{\text{i.i.d.}}{\sim} \text{Gamma} (\delta, \delta) \end{cases} \quad k = 1, \dots, K. \tag{9}$$

Equation (9) models a prior opinion of homogeneity among donors. Our choice corresponds to a discretized approximation of a very flexible non-parametric prior for the cumulative hazard, that is the gamma process prior (Kalbfleisch 1978). We have:

$$E(u_i(t) \mid \lambda_{01}, \dots, \lambda_{0K}, c) = \sum_{k=1}^K E(u_{ik} \mid \lambda_{0k}, c) \mathbf{1}_{(a_{k-1}, a_k]}(t) = \sum_{k=1}^K \lambda_{0k} \mathbf{1}_{(a_{k-1}, a_k]}(t), \tag{10}$$

so that:

$$E(u_i(t)) = \sum_{k=1}^K E(\lambda_{0k}) \mathbf{1}_{(a_{k-1}, a_k]}(t) = 1. \tag{11}$$

The hyperparameters  $c$  and  $\delta$  quantify the uncertainty on the steps  $u_{ik}$  and the pairwise correlation between frailties. In particular, they measure how the prior of the steps  $\{u_{ik}\}$  widenspreads around its mean  $E(u_i(t))$  because:

$$\text{Var}(u_i(t) \mid \lambda_{01}, \dots, \lambda_{0K}, c) = \frac{\sum_{k=1}^K \lambda_{0k} \mathbf{1}_{(a_{k-1}, a_k]}(t)}{c} \tag{12}$$

and hence:

$$\text{Var}(u_{ik}) = \frac{1}{c(a_k - a_{k-1})} + \frac{1}{\delta}.$$

In addition, the correlation  $\rho(u_i(s), u_j(t))$  between  $u_i(s)$  and  $u_j(t)$  changes with  $c$  and  $\delta$  as:

$$\rho(u_i(s), u_j(t)) = \frac{c(a_k - a_{k-1})}{c(a_k - a_{k-1}) + \delta}$$

for all  $s, t \in (a_{k-1}, a_k]$  and  $k = 1, \dots, K$ , while  $\rho(u_i(s), u_j(t)) = 0$  otherwise.

Equation (10) provides an immediate interpretation of the step function  $\lambda_0(t) := \sum_{k=1}^K \lambda_{0k} \mathbf{1}_{(a_{k-1}, a_k]}(t)$  as the baseline intensity function. In the Bayesian framework, the hyperparameters  $\{\lambda_{01}, \dots, \lambda_{0K}\}$  form a piecewise constant baseline intensity function. As a consequence, the posterior mean  $\hat{\lambda}_0(t)$  of  $\lambda_0(t)$ , given by

$$\hat{\lambda}_0(t) = \sum_{k=1}^K E(\lambda_{0k} | \text{data}) \mathbf{1}_{(a_{k-1}, a_k]}(t),$$

is a Bayesian estimate of the baseline intensity function  $\lambda_0(t)$ . The cut-points of the step functions  $u_i(t)$ 's over the time window have been taken equally spaced. The marginal prior (9) turns out to be a convenient choice, since in this case  $\sum_{k=1}^K \lambda_{0k} \mathbf{1}_{(a_{k-1}, a_k]}(t)$  can be interpreted as the *centering* hazard for all donors. Finally, in accordance with the scale-invariance of the gamma prior of this parameterization, the parameter  $v_{ik}$ , defined by

$$v_{ik} = \frac{u_{ik}}{\lambda_{0k}}, \tag{13}$$

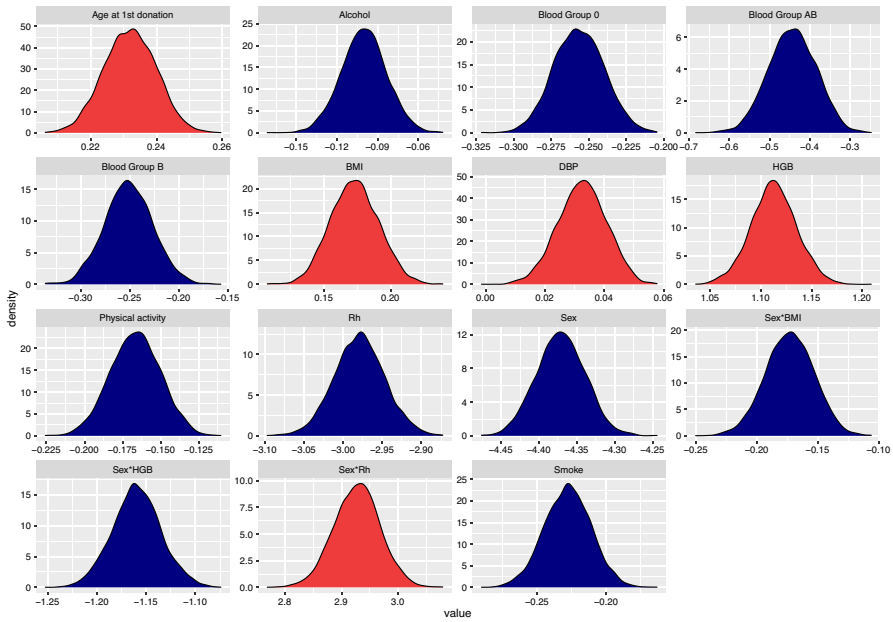
represents the specific random effect of donor  $i$  at time  $t \in (a_{k-1}, a_k]$  for  $k = 1, \dots, K$ . The terms  $\{v_{ik}\}_k$  have a multiplicative effect on the intensity function of donor  $i$ , so that  $v_{ik} > 1$  indicates more propensity to experience a donation and vice versa, with the same covariate values.

### 5 Posterior inference

We compute the posterior distribution of the Bayesian model (7)–(9) using a Markov chain Monte Carlo (MCMC) algorithm. In particular, we employ the Hamiltonian Monte Carlo (HMC) through the software platform Stan (Stan Development Team 2020). The values of the hyperparameters in (8)–(9) are set to  $\delta = 2$ ,  $c = 0.01$  and  $\sigma_0^2 = 10^4$ . We run two HMC chains of 5000 iterations, each one with a warmup of 3000 iterations, for a final sample of 4000 in total; MCMC convergence diagnostics available in the R package CODA show that convergence holds for all parameters (Plummer et al. 2006).

The posterior densities of the regression parameters in  $\beta$  for the Bayesian model with all 17 covariates including interactions are highly concentrated around 0. Therefore, we have discarded SBP and HR from the model, and the new model has been refitted. Figure 4 shows the marginal posterior densities of the 15 remaining covariates, in red if the factor increases the intensity of recurrence and in blue vice versa.

A posteriori,  $\beta_{\text{SEX}}$  is concentrated on negative values, suggesting that male donors have a smaller recurrence intensity than female donors. Individuals with blood groups O, B, or AB exhibit a lower recurrence intensity compared to the reference blood group A. Similarly, donors with RH<sup>+</sup> factor have a smaller recurrence intensity than donors with the reference RH<sup>-</sup>. This suggests that, since RH<sup>-</sup> factor is much rarer than RH<sup>+</sup> in the dataset but also in the Italian population, donors with negative Rh factor are more involved and regularly donate. Moreover, male donors with RH<sup>-</sup> factor donate more frequently than female donors with RH<sup>-</sup>. Smokers and drinkers tend to donate less than non-smokers and non-drinkers, while non-active donors (PHYSICAL\_ACTIVITY = 0) are more likely to donate. Finally, HGB, DBP, age and BMI have a positive effect on the recurrence intensity. However,



**Fig. 4** Marginal posterior densities of the regression parameters and interactions in  $\beta$  after discarding SBP and HR. Blue and red colors discriminate between negative and positive effects of the factors on the intensity function, respectively

**Table 3** Posterior mean, standard deviations and quantile of the baseline parameters  $\lambda_{0k}$

Parameter	Mean	Standard deviation	$q_{0.025}$	$q_{0.5}$	$q_{0.975}$
$\lambda_{01}$	5.50	0.12	5.26	5.50	5.74
$\lambda_{02}$	0.87	0.02	0.84	0.87	0.90
$\lambda_{03}$	0.60	0.01	0.57	0.60	0.63
$\lambda_{04}$	0.50	0.01	0.47	0.50	0.52
$\lambda_{05}$	0.41	0.01	0.39	0.41	0.43
$\lambda_{06}$	0.42	0.01	0.40	0.42	0.45
$\lambda_{07}$	0.39	0.01	0.36	0.39	0.42
$\lambda_{08}$	0.47	0.02	0.44	0.47	0.50
$\lambda_{09}$	0.56	0.02	0.51	0.55	0.60
$\lambda_{010}$	1.01	0.05	0.92	1.01	1.13

note that the effects of HGB and BMI are stronger for female donors than for male donors, since the marginal posteriors of  $\beta_{SEX*HGB}$  and  $\beta_{SEX*BMI}$  are concentrated on negative values.

Table 3 reports posterior summaries of the baseline parameters  $\lambda_{0k}$ . The estimation of the piecewise baseline intensity function  $\lambda_0(t)$  suggests that donors are more likely to donate in their first year and that this propensity tends to decrease over time.

Figure 5 displays the posterior means of the individual random effect functions  $\{v_{ik}\}_{k=1,\dots,K}$  defined in Eq. (13) for 10 randomly selected male and female donors. These plots show the random heterogeneity of the individuals in the sample, and express the variability that the other parameters of the model cannot quantify. From the figure we also see that the values of frailties are typically greater for men than for women. This is also confirmed if we plot the posterior means of more individual frailties, which are not included here. This is an expected result, because of the Italian donation rules for which men can donate more often than women.

We have reported above the posterior inference when hyperparameter  $c$  is set to 0.01. Since this hyperparameter controls both the variance of each  $u_{ik}$  and the correlation between  $u_{ik}$  and  $u_{jk}$ , we have also fitted the model for  $c = 1$  and  $c = 2$ . The posterior estimates turned out to be robust and, for this reason, they are not reported here.

### 6 Profiling and prediction

Any blood collection center must deal with donor profiling and donation prediction.

*Profiling* is a key task to carry out effective campaigns to acquire further donors. Indeed, donor recruitment campaigns should be directed towards individuals whose characteristics could guarantee a high frequency of donation and continuity over the years. In this light, the proposed approach allows for predicting future donation patterns of each possible donor profile. More precisely, we can compute the posterior predictive distribution of potential novel donors, identified by representative values of the vector of covariates. As a result, the collection center will be able to

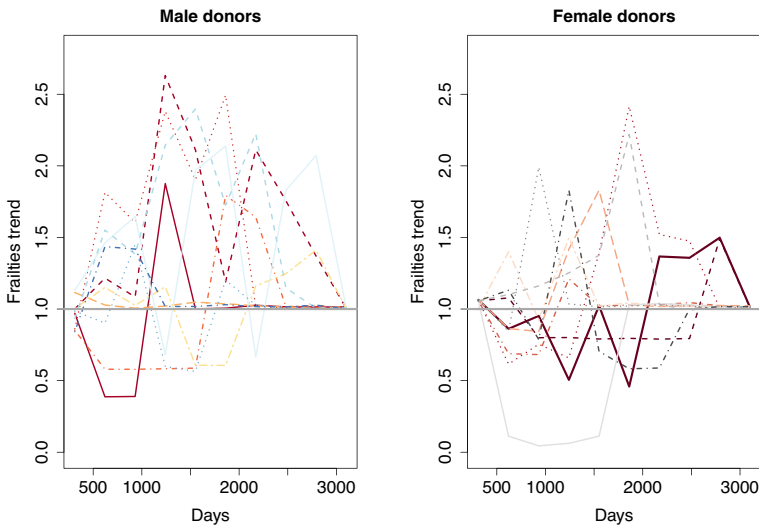


Fig. 5 Posterior means of the random effect functions  $v_{i1}, \dots, v_{i10}$  for ten randomly selected male (left) and female (right) donors

appropriately choose the target of a recruitment campaign, directing it towards the most promising profiles, e.g., in high schools (very young donors) rather than in companies (older donors) or vice versa.

*Prediction* refers to the estimation of the blood units that will be donated in total and/or for each blood type over a given time horizon. This is required for the internal organization, to size the capacity and personnel needed on any day of the time horizon, and to predict the achievement of production targets for each blood type. In particular, for the internal organization, the time horizon is usually one week and there is no need to distinguish between different blood types. Instead, the production targets are set for a longer time horizon (e.g., one month) and for each blood type individually. In our framework, information can be provided in terms of the average number of donors over the time horizon, because each donation corresponds to a single blood unit and a donor can donate at most once a week or in a month.

## 6.1 Profiling

The aim of profiling is to identify donor profiles who donate more frequently, discovering the characteristics that significantly yield shorter waiting times before the next donation. With this aim, we compute the posterior predictive probability that a *new* donor  $l$  identified by covariates  $\mathbf{x}_l$  will donate after  $t$  days from their first donation. In particular, the posterior predictive probability that the *new* donor  $l$  will donate after at least  $t$  days from the first donation can be computed as the MCMC mean of  $P(W_{l1} > t | T_{l0} = 0, \text{parameters})$ , where we assume that the first donation of the new donor  $l$  is made at time  $T_{l0} = 0$ , so that  $W_{l1} = T_{l1} - T_{l0}$  is the waiting time of the first recurrence (corresponding to the second donation). This probability can be derived from Corollary 1 of Chapter 1 in Cook and Lawless (2007). We let  $t$  vary in the first three months in which donor  $l$  is allowed to donate, i.e.,  $t \in (\Phi_l + 1, \Phi_l + 90)$  where  $\Phi_l = 85$  days for men and 150 for women. In this way, only  $u_{l1}$ , which covers the first 310 days of the donation process, is needed. We obtain:

$$\begin{aligned} P(W_{l1} > t | T_{l0} = 0, \text{parameters}) &= \exp \left\{ - \int_0^t \lambda_l(s) ds \right\} = \exp \left\{ - \int_0^t u_{l1} e^{\mathbf{x}_l'(s)\beta} Y_l(s) ds \right\} \\ &= \begin{cases} 1 & \text{if } t \leq \Phi_l \\ \exp \{ -u_{l1} e^{\mathbf{x}_l'\beta} (t - \Phi_l) \} & \text{if } t > \Phi_l \end{cases} \end{aligned}$$

where we have assumed that the time-varying covariate vector  $\mathbf{x}_l(s)$  is constant between the two successive donations considered, and we have denoted this value as  $\mathbf{x}_l$ .

A few relevant profiles have been selected and reported in Table 4 to provide a concrete example. Profile 0 refers to a healthy young man with an age at first donation equal to 25 years. Profile 1 describes a healthy young woman, 20 years old, with the rest of the covariates as in Profile 0. Profiles 2 and 3 refer to healthy middle-aged donors (a male donor and a female donor). Profiles 4–7 refer to middle-aged men and women with unhealthy lifestyle habits. Finally, Profiles 8 and 9 correspond to Profiles 0 and 1, but with the less common blood type AB and negative Rh factor. HGB and DBP of each profile are set to the sample mean of the same-sex and

**Table 4** Selected profiles of potential donors

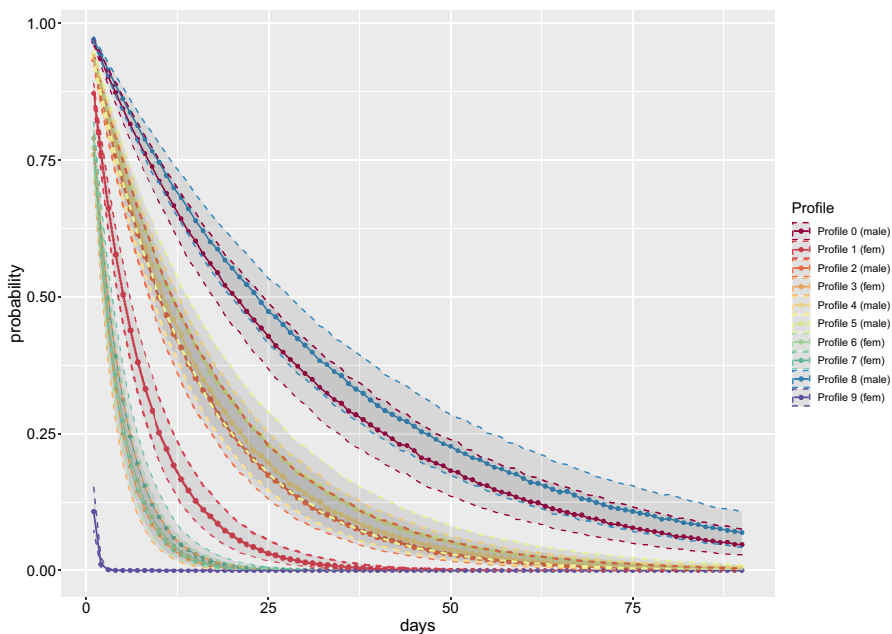
	Profiles									
	0	1	2	3	4	5	6	7	8	9
Sex	Male	Female		Female		Female		Female		Female
Age at first donation	25	20	55	50	55	55	50	50	50	20
Weight	75 kg	65 kg	85 kg	65 kg	95 kg	95 kg	85 kg	85 kg	85 kg	65 kg
Height	1.75 m	1.63 m		1.63 m			1.63 m	1.63 m	1.63 m	1.63 m
Range BMI	Healthy weight				Overweight	Overweight	Overweight	Overweight	Overweight	Overweight
Smoke	No				Yes	Yes	Yes	Yes	Yes	Yes
Alcohol	No				No	Yes	Yes	Yes	Yes	Yes
Physical activity	Yes				No	No	No	No	No	No
Blood group	0						No			AB
Rh factor	+									AB

Profiles 2 and 3 represent elder donors than Profiles 0 and 1; Profiles 4–7 amount to unhealthy lifestyle habits; Profiles 8 and 9 consider different blood groups and Rh factors than Profiles 0 and 1, respectively. Empty cells denote the same value as in the reference Profile 0

age-matched donor subset. Finally, values of all covariates have been kept constant during the analyzed 90-day period in each profile.

Figure 6 shows the posterior mean and 95% posterior credible bounds of  $P(W_{11} > t | T_{10} = 0, \mathbf{x}_t, \text{parameters})$ , for 10 simulated *new* donors ( $l = 0, 1, \dots, 9$ ). The faster these graphs decrease, the more the associated donors are more likely to donate as soon as they are allowed, which indicates more productive donors.

Profile 9 is the most productive, which refers to a woman who starts to donate at 20 years old, with a healthy weight, active life, no smoking, no alcohol and rare blood type AB with negative Rh factor. This agrees with comments made in Sect. 3 about exploratory data analysis, namely women donate more often than men, subject to legal obligations. Furthermore, the posterior analysis (Sect. 5) showed that  $\beta_{\text{SEX}} < 0$  for women, providing evidence that women have a higher recurrence intensity compared to men. The next three best profiles (Profiles 3, 6 and 7) all represent middle-aged female donors. The posterior predictive probability that they will donate after at least  $t$  days from the first donation declines very rapidly in the first three weeks after  $\Phi_l$  and later approaches to zero. On the contrary, Profiles 0 and 8, both referring to young men, show the lowest propensity to donate as the associated posterior predictive survival probability decreases more slowly. The trajectories of male donors are clearly grouped by age: one group corresponds to middle-aged Profiles 2, 4 and 5, and the other to young Profiles 0 and 8. As for blood group and Rh factor, we observe from Fig. 6 that the effect of GROUP\_AB together with negative Rh factor is significant for women only; in particular, note how Profiles 1 and



**Fig. 6** Bayesian posterior prediction of  $P(W_1 > t | T_0 = 0, \mathbf{x}, \text{parameters})$  in the first 90 days in which each profile is allowed to donate after his/her first donation. 95% credible bounds are added as dashed lines

9, both corresponding to the same female profile but with different blood group and Rh factor, differ so much. At the end of the three months, all profiles but Profile 2 show a posterior predictive survival probability that approaches to zero, meaning that they are highly likely to have made the second donation within the analyzed time window. To conclude, among these profiles, we suggest AVIS Milan to target new recruitment campaigns for middle-aged men and young women.

### 6.2 Prediction

The estimate of the daily production of blood units can be obtained from the posterior inference, by combining the predictions of all donors who might donate within a given horizon. Since our model does not account for the arrival of novel donors but only for the recurrent donations process, we can only estimate the number of blood units from the donors already in the sample. This represents a lower bound for the estimate of the total number of blood units that can be collected.

The number of donations from the donors in the dataset that will take place in the next  $\tau$  days after June 30th, 2018 (last day of observation in the database) can be expressed as:

$$N(\tau) = \sum_{i=1}^I \mathbf{1}(N_i(c_i + \tau) - N_i(c_i) = 1 | \boldsymbol{\beta}, \mathbf{u}, \lambda_0), \tag{14}$$

with (conditional) expected value:

$$\mu(\tau) = E(N(\tau) | \boldsymbol{\beta}, \mathbf{u}, \lambda_0) = \sum_{i=1}^I P(N_i(c_i + \tau) - N_i(c_i) = 1 | \boldsymbol{\beta}, \mathbf{u}, \lambda_0). \tag{15}$$

A Bayesian estimator  $\hat{\mu}(\tau)$  of  $\mu(\tau)$  is given by its posterior mean:

$$\hat{\mu}(\tau) = E(\mu(\tau) | \text{data}) = \sum_{i=1}^I E(P(N_i(c_i + \tau) - N_i(c_i) = 1 | \text{data})). \tag{16}$$

Such conditional expected value  $\mu(\tau)$  should be at least equal to the production target. On the contrary, an average number of donations higher than the target is not considered as a problem, because extra units can be stored or distributed to neighboring facilities. In addition, potential recurrent donors could be asked to delay the next donation to periods of underproduction.

The at-risk indicator  $Y_i(s)$  of donor  $i = 1, \dots, I$  at day  $s \in (c_i, c_i + \tau]$  turns out to be:

$$Y_i(s) = \begin{cases} 1 & \text{if } s - t_{in_i} \geq \Phi_i \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

where  $t_{in_i}$  is the day of the last donation for donor  $i$ , and  $\Phi_i$  is again equal to 85 for men and 150 for women. In the light of those remarks, it follows from Eq. (6) that:

$$P(N_i(c_i + \tau) - N_i(c_i) = 1 | \beta, \mathbf{u}, \lambda_0) = 1 - \exp \left\{ - \int_{c_i}^{c_i + \tau} u_i(s) e^{x_i' \beta} Y_i(s) ds \right\} \quad (18)$$

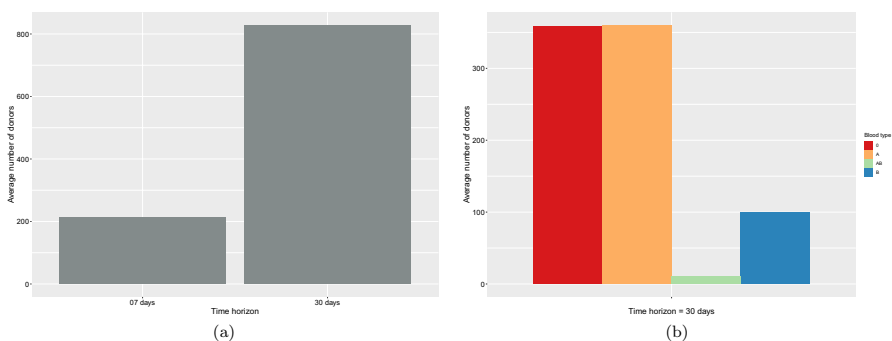
with Bayesian estimator  $\hat{p}_i := E(P(N_i(c_i + \tau) - N_i(c_i) = 1) | \text{data})$  given by:

$$\hat{p}_i = \begin{cases} 1 - E \left( \exp \left\{ - \int_{c_i}^{c_i + \tau} u_i(s) e^{x_i' \beta} ds \right\} | \text{data} \right) & \text{if } t_{in_i} + \Phi_i \leq c_i \\ 1 - E \left( \exp \left\{ - \int_{t_{in_i} + \Phi_i}^{c_i + \tau} u_i(s) e^{x_i' \beta} ds \right\} | \text{data} \right) & \text{if } c_i \leq t_{in_i} + \Phi_i \leq c_i + \tau \\ 0 & \text{if } t_{in_i} + \Phi_i \geq c_i + \tau. \end{cases} \quad (19)$$

Let us now focus on a time horizon of either a week ( $\tau = 7$ ) or a month ( $\tau = 30$ ) after June, 30<sup>th</sup> 2018, which are of practical importance. Figure 7a reports the estimate of  $\mu(7)$  and  $\mu(30)$  and shows consistent values, since  $\mu(30)$  is approximately equal to four times  $\mu(7)$ . Figure 7b shows the estimate of  $\mu(30)$  per blood type, in which the proportion between groups is consistent, with the largest number of blood units for groups 0 and A, followed by group B. Compared to AVIS Milan practice, the observed monthly number of blood units produced in 2018 was approximately 1500. However, note that this number includes not only the donations from the donors in our dataset, but also from non-recurrent donors and from recurrent donors already active before January 1st, 2010. In particular, the donations from the donors in our dataset (for which we make a prediction) represent about 60% of all donations recorded by AVIS Milan in the time window we focus on, corresponding to about 30 donations recorded per day. Therefore, our  $\mu(7)$  and  $\mu(30)$  are coherent.

## 7 Conclusions

Uncertainty is a major issue when managing health care facilities and affects the performance of the service. This is particularly true when considering blood collection centers, which merge the features of a service provider and those of a



**Fig. 7** Estimated number of donors  $\mu(\tau)$  for the next week  $\tau = 7$  and the next month  $\tau = 30$ , after June 30 th, 2018 (a); estimated number of donors  $\mu(30)$  grouped by blood group (b)

production system. The ability to predict the times of donations is a key point to guarantee an adequate production of blood units and at the same time to properly manage the resource at the blood collection center. We focus our attention on two points in particular, profiling and prediction, and we propose a statistical prediction model that quantitatively supports these tasks.

As for profiling, the use of the proposed model can make the process of acquiring further donors more efficient, directing interventions towards donors who, once enrolled, will donate as often as they are allowed by the law. Our model indicates the most productive donor profiles among a set of possible alternatives. As for prediction, our model is able to predict the amount of donations in the next days. However, our approach considers only recurrent donors entered after the initial date of the time window, and provides estimates only for them. Therefore, when evaluating the number of donations in a time horizon, those from non-recurrent donors and recurrent donors already active before the time window should be added. To predict the number of non-recurrent donors, we could consider a model describing the arrivals of novel donors; however, according to AVIS Milan staff, in recent years the amount of donations from novel donors seems negligible with respect to the others. Note that we have excluded recurrent donors already active before the time window since their inclusion would require left-censoring in the likelihood in addition to right-censoring. This extension is among our future works.

General indications can also be drawn for the case of AVIS Milan, which may result in useful qualitative knowledge, supported by quantitative analyses, to be provided to the staff. In particular, our analysis highlights a decreasing trend of the baseline intensity function. It also identifies individual features (sex, smoking habits, alcohol consumption, physical activity, BMI, Rh factor, blood group, age at first donation, hemoglobin and minimum pressure) that most influence the intensity function and, hence, determine donors' personal propensity to donate. Also, the interactions between sex and hemoglobin, Rh factor and BMI are found to be significant in differentiating donors' behavior.

Our method can be immediately applied to the AVIS Milan case and also to other blood collection centers. In the latter case, it is enough to compute the posterior distribution of the parameters, given the different data. Then, the results of similar analyses on other databases can provide helpful information for centers. In conclusion, we have proposed a model that has proven to be an effective solution for profiling and prediction needs, and that can be immediately used in any blood collection center.

**Acknowledgements** The authors would like to thank Ilaria Martinelli for her support in extracting the data and writing the code for posterior simulations. The first and third authors acknowledge the support by MUR, Grant Dipartimento di Eccellenza 2023–2027.

**Funding** Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Addis B, Carello G, Grosso A, Lanzarone E, Mattia S, Tànfani E (2015) Handling uncertainty in health care management using the cardinality-constrained approach: advantages and remarks. *Oper Res Health Care* 4:1–4
- Aldamiz-Echevarria C, Aguirre-Garcia MS (2014) A behavior model for blood donors and marketing strategies to retain and attract them. *Rev Lat Am Enfermagem* 22(3):467–475
- Alkahtani AS, Jilani M (2019) Predicting return donor and analyzing blood donation time series using data mining techniques. *Int J Adv Comput Sci Appl*. <https://doi.org/10.14569/ijacsa.2019.0100816>
- Ayer T, Zhang C, Zeng C, White CC III, Joseph VR, Deck M, Lee K, Moroney D, Ozkaynak Z (2018) American red cross uses analytics-based methods to improve blood-collection operations. *Interfaces* 48(1):24–34
- Ayer T, Zhang C, Zeng C, White CC III, Joseph VR (2019) Analysis and improvement of blood collection operations. *Manuf Serv Oper Manag* 21(1):29–46
- Baş S, Carello G, Lanzarone E, Ocak Z, Yalçındağ S (2016) Management of blood donation system: literature review and research perspectives. *Health care systems engineering for scientists and practitioners*. Springer, Cham, pp 121–132
- Baş S, Carello G, Lanzarone E, Yalçındağ S (2018) An appointment scheduling framework to balance the production of blood units from donation. *Eur J Oper Res* 265(3):1124–1143
- Baş Güre S, Carello G, Lanzarone E, Yalçındağ S (2018) Unaddressed problems and research perspectives in scheduling blood collection from donors. *Prod Plan Control* 29(1):84–90
- Beliën J, Forcé H (2012) Supply chain management of blood products: a literature review. *Eur J Oper Res* 217(1):1–16
- Bischoff F, Koch MdC, Rodrigues PP (2019) Predicting blood donations in a tertiary care center using time series forecasting. *ICT for health science research*. IOS Press, Amsterdam, pp 135–139
- Boonyanusith W, Jittamai P (2012) Blood donor classification using neural network and decision tree techniques. In: *Proceedings of the world congress on engineering and computer science*, pp 499–503
- van Buuren S, Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in R. <https://www.jstatsoft.org/v45/i03/>
- Christensen R, Johnson W, Branscum A, Hanson TE (2010) *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC Press, Boca Raton
- Cook RJ, Lawless J (2007) *The statistical analysis of recurrent events*. Springer, Berlin
- Darwiche M, Feuilloley M, Bousaleh G, Schang D (2010) Prediction of blood transfusion donation. In: *2010 fourth international conference on research challenges in information science (RCIS)*. IEEE, pp 51–56
- Doneda M, Yalçındağ S, Lanzarone E (2023) A three-stage matheuristic for home blood donation appointment reservation and collection routing. *Flex Serv Manuf J*. <https://doi.org/10.1007/s10696-023-09518-6>
- Hamdan B, Diabat A (2019) A two-stage multi-echelon stochastic blood supply chain problem. *Comput Oper Res* 101:130–143
- Jabbarzadeh A, Fahimnia B, Seuring S (2014) Dynamic supply chain network design for the supply of blood in disasters: a robust model with real world application. *Transp Res Part E Logist Transp Rev* 70:225–244
- Kalbfleisch JD (1978) Non-parametric Bayesian analysis of survival time data. *J R Stat Soc Ser B (Methodological)* 40(2):214–221

- Kassie A, Birara S (2021) Practice of blood donation and associated factors among adults of Gondar city, northwest Ethiopia: Bayesian analysis approach [retraction]. *J Blood Med* 12:85–86
- Kauten C, Gupta A, Qin X, Richey G (2021) Predicting blood donors using machine learning techniques. *Inf Syst Front* 24:1–16
- Khalid NSC, Burhanuddin M, Ahmad A, Ghani M (2013) Classification techniques in blood donors sector—a survey. In: E-proceeding of software engineering postgraduates workshop (SEPoW)
- Khalilinezhad M, Dellepiane S, Abedi F, Vernazza G (2014) Extracting hidden patterns in blood donor database using association rule mining. In: Proceedings of the European conference on data mining 2014 and international conferences on intelligent systems and agents 2014 and theory and practice in modern computing 2014, pp 12–20
- Kircic P, Aktas S, Sevinc B (2020) Analyzing blood donation probabilities and number of possible donors. In: 2020 international congress on human-computer interaction, optimization and robotic applications (HORA). IEEE, pp 1–4
- Klein JP, Moeschberger ML (2003) Survival analysis: techniques for censored and truncated data. Springer, Berlin
- Lanzarone E, Yalçındağ S (2019) Uncertainty in the blood donation appointment scheduling: key factors and research perspectives. In: International conference on health care systems engineering, Springer, pp 293–304
- Mohammadi T, Kheiri S, Sedehi M (2016) Analysis of blood transfusion data using bivariate zero-inflated Poisson model: a Bayesian approach. *Comput Math Methods Med*. <https://doi.org/10.1155/2016/7878325>
- Osorio AF, Brailsford SC, Smith HK (2015) A structured review of quantitative models in the blood supply chain: a taxonomic framework for decision-making. *Int J Prod Res* 53(24):7191–7212
- Plummer M, Best N, Cowles K, Vines K (2006) Coda: convergence diagnosis and output analysis for MCMC. *R News* 6(1):7–11
- Rabbani M, Aghabegloo M, Farrokhi-Asl H (2017) Solving a bi-objective mathematical programming model for bloodmobiles location routing problem. *Int J Ind Eng Comput* 8(1):19–32
- Ramachandran P, Girija N, Bhuvanewari T (2011) Classifying blood donors using data mining techniques. *Int J Comput Sci Eng Technol* 1(1):10–13
- Ramezani R, Behboodi Z (2017) Blood supply chain network design under uncertainties in supply and demand considering social aspects. *Transp Res Part E Logist Transp Rev* 104:69–82
- Santhanam T, Sundaram S (2010) Application of cart algorithm in blood donors classification. *J Comput Sci* 6(5):548
- Shashikala B, Pushpalatha M, Vijaya B (2019) Machine learning approaches for potential blood donors prediction. *Emerging research in electronics, computer science and technology*. Springer, Berlin, pp 483–491
- Stan Development Team (2020) RStan: the R interface to Stan. R package version 2.21.2. <http://mc-stan.org/>
- Sundaram S, Santhanam T (2011) A comparison of blood donor classification data mining models. *J Theor Appl Inf Technol* 30(2):98–101
- Tavakol N, Kheiri S, Sedehi M (2016) Analysis of the factors affecting the interval between blood donations using log-normal hazard model with gamma correlated frailty. *J Res Health Sci* 16(2):76
- Testik MC, Ozkaya BY, Aksu S, Ozcebe OI (2012) Discovering blood donor arrival patterns using data mining: a method to investigate service quality at blood centers. *J Med Syst* 36(2):579–594
- World Health Organization (2012) Blood donor selection: guidelines on assessing donor suitability for blood donation. World Health Organization
- Yalçındağ S, Güre SB, Carello G, Lanzarone E (2020) A stochastic risk-averse framework for blood donation appointment scheduling under uncertain donor arrivals. *Health Care Manag Sci* 23(4):535–555
- Zahiri B, Torabi S, Mousazadeh M, Mansouri S (2015) Blood collection management: methodology and application. *Appl Math Model* 39(23–24):7680–7696

**Ilenia Epifani** received the Ph.D. in Probability and Statistics from the University of Trento, Italy, in 1999. She is an associate professor of Statistics at the Politecnico di Milano, Italy. Her recent research interests are in spatial Bayesian models and their statistical applications in health care management, human geography (dynamic of urban population), environmental science (air quality, drinking water disinfection, interplay between natural resources and social conflict dynamic).

**Ettore Lanzarone** received the Ph.D. in Bioengineering from the Politecnico di Milano, Italy, in 2008. He is an associate professor at the Department of Management, Information and Production Engineering of the University of Bergamo, Italy. He is also a research associate at the Institute for Applied Mathematics and Information Technology (IMATI) of the National Research Council (CNR), Italy, where he worked as a researcher from 2011 to 2020. He is a member of the CIRRELT (Centre Interuniversitaire de Recherche sur les Reseaux d'Entreprise, la Logistique et le Transport), Canada. His research activities focus on bioengineering, optimization and operations research, and stochastic models. Applications of the last two topics include healthcare (e.g., home care, blood donation system, ambulance tracking) and manufacturing (scheduling). He recently chaired the 48th annual meeting of the EURO working group on Operational Research Applied to Health Services (ORAHS 2022).

**Alessandra Guglielmi** received the Ph.D. in Mathematics from the University of Milano, Italy, in 1997. She is a full professor of Statistics at the Politecnico di Milano, Italy. Her recent research interests are in Bayesian mixture models for density estimation and clustering. Her work also concerns statistical applications in medicine, health care management, air quality modeling, and Bayesian optimization for cloud systems and big data applications.