Transfusion Medicine, 2000, **10,** 181–192

Predicting a donor's likelihood of donating within a preselected time interval

W. A. Flegel,* W. Besenfelder† and F. F. Wagner* **Department of Transfusion Medicine, University of Ulm, DRK (German Red Cross)-Blood Service Baden-Württemberg, Institute Ulm, Ulm, Germany, and †Department Biometrics and Medical Documentation, University of Ulm, Ulm, Germany*

Received 1 October 1999; accepted for publication 8 May 2000

SUMMARY. The procurement of some advanced blood components, like quarantined plasma units, depends critically on retesting the donor within a fixed time frame. For health care systems, such as that in Germany, with mandatory retesting of donors before plasma release, the reliable identification of donors who are more likely to return in time has an immense practical implication, because their blood components could be preferably selected for quarantine purposes. The donation histories of about 760 000 donors with 4910 000 donation attempts were analysed. We developed a logistic regression model to calculate a probability of donation, $p(D_{ts-te})$, within a preselected time frame (t_s–t_e). The donation history was compounded in a score and shown to be very useful for determining $p(D_{ts-te})$. A logistic regression model was developed with score and donor status as parameters; different regression coefficients applied to first-time-donors (ftd) and to repeat donors (intercept, int, and score factor, scf). This model allowed us to determine the probability of donation, $p(D_{ts-te})$, within a preselected time interval, e.g. 6–9 months after an index donation. The $p(D_{ts-te})$ can be calculated for any donor of blood services. The $p(D_{170-275\,days})$ ranged from about 22% to 86% for any index donation in 1996/97. First-time donors had a $p(D_{170-275\,days})$ of 33% and were more likely to return within the time interval than certain subsets of repeat donors who can be defined by our model. We provided a technical procedure to increase the rate of plasma unit release after quarantine storage and showed the usefulness of our procedure for blood component management, if quarantine storage is required. By applying the model to our current plasma quarantine programme we could retrieve about 30% more units, which would represent about 30 000 units per year, without incurring additional costs. General implications for blood collection, like planning blood drives, were discussed. The whole demand of a health care system for single plasma units may be met by quarantine plasma and their cost-efficiency can be improved.

Key words: blood collection, blood donation, blood donor, cost containment, cryopreserved red cell units, first-time donor, quarantine, quarantined plasma units, repeat donor, time interval.

Advanced blood components, like quarantined plasma and cryopreserved red cell units, can be stored for months before being transfused. To further enhance the virus safety of such blood components, a mandatory retesting of the donors for infectious disease markers has been introduced in the mid 1990s into transfusion practice of several countries, like Germany, for which our observation will be important. Since then the supply of these blood components has depended critically on the co-operation of the donors within a rather tight time frame after their index donation. Because less than half of all whole blood donations are required for sufficient supply with quarantine plasma units, a selection of donations suitable for quarantine became feasible. The reliable identification of donors who can be expected to return within the relevant time frame would have an immense practical implication for health care systems with quarantine requirements. Plasma that does not meet quarantine criteria may only be used for plasma fractionation purposes.

Recently, Whyte (1999) analysed about 560 000

Correspondence: Willy A. Flegel, Priv.-Doz. Dr med., Abteilung Transfusionsmedizin, Universitätsklinikum Ulm, and DRK-Blutspendedienst Baden-Württemberg, Institut Ulm; Helmholtzstrasse 10; D-89081 Ulm, Germany. Tel.: +49 731 150 600; fax: +49 731 150 602; e-mail: waf@ucsd.edu

donation intervals and found the aggregate donor behaviour to be very predictable. With a simple regression analysis, the attendance of repeat donors and of first-time donors was shown to be related.

James & Matthews (1993, 1996) developed a model based on a 'donation cycle' by analysing about 180 000 blood group O donors and 610 000 donation attempts. To predict the time that will elapse since the previous donation attempt, they applied a Cox regression model (Cox & Oakes, 1984) as used in standard time-to-outcome methods, which yield survival curves and relative risk estimates (James & Matthews, 1996). They showed some influence of sex, Rhesus blood group and rare donor status on donor return. They also analysed the donor history for up to four previous donation cycles and concluded that the elapsed time since a previous index donation affected the likelihood of a subsequent donation significantly (James & Matthews, 1996).

Many major blood donation services like ours run blood drives at temporary countryside locations. Donors usually give blood at their village location or nearby, where blood drives by our blood service are offered 2–4 four times a year. The usual minimum interval between blood drives is 82 days. Hence, it was possible to define a dichotomous variable, successful attending or not attending an offered blood drive at particular locations. In contrast to previous approaches (James & Matthews, 1996) our model obviated the need to calculate time intervals between donations and their variation. We analysed the utility of a donor's donation history to calculate a 'probability of donation', p(D), within a preselected time frame. Other donor demographic factors were also analysed, but found to be much less useful for this purpose. Mathematically, our approach allowed us to apply a logistic regression model (Hosmer & Lemeshow, 1989) and obviated the need for the more complex Cox regression. The application of the model to our current plasma quarantine program and the considerable practical benefits for program management were demonstrated. We discussed the model's broader implications for blood procurement.

MATERIALS AND METHODS

Blood donor database and descriptive statistics

Since 2 May 1985, a central electronic database for blood donation has been maintained at the DRK-Blutspendedienst Baden-Württemberg, Germany (Wagner *et al*., 1995) comprising all donation-related information including location and laboratory results. On 5 November 1998, records for 763 401 donors and 4915 777 donations at 1150 different locations were available for the current study. Our data sets represented more than 7% of all residents in the *Land* (state of) Baden-Württemberg and more than 80% of all donations in that area. All donor data were reviewed and 49 219 donors (6·45%) were excluded from evaluation for reasons such as inconsistent data or positive infectious disease markers. Donations at the three permanent blood donation sites of the blood donation service in the cities of Ulm, Mannheim and Baden-Baden were also excluded. The examined data thus accurately represented the donation patterns at temporary donation sites.

Since 1 January 1995, we observed a 6-month quarantine before plasma units were released, a procedure made mandatory on 1 July 1995 by the German regulatory authority, the Paul-Ehrlich-Institut, Langen. Starting in mid 1994, all donors were informed about the quarantine requirements and asked to return for retesting purposes. The donors were notified by postcards once a blood drive was offered at the location of their most recent donation.

Descriptive donor statistics were performed by standard methods. The influence of various demographic characteristics on donor return within 170–275 days after an index donation was calculated in a univariate analysis. R^2, the coefficient of determination, was calculated as described by Nagelkerke (1991).

Development of a logistic regression model

To determine relevant demographic factors for returning to a subsequent blood drive and to develop a model for predicting the probability of donation within a preselected time frame, we evaluated the subset of 210 786 donors who donated blood between 1 October 1995 and 30 September 1996. The donors were randomly assigned to one of three groups. With the first group, a logistic regression model was constructed, which was validated with the second group as an independent sample. Because we considered this model satisfactory, a further modelling step was not required, for which we would have spared the third group as another independent validation sample.

Donor score. A donor score was defined as in eqn 1 and, based on the donor's donation record, can be determined for any donor at our blood donation service.

If a donor had donated blood at the latest blood drive at her or his location before a current donation, i.e. at the most recent previous opportunity to donate blood, a 1 was added to the 'credit'. If the donor donated at the next most recent occasion, a 0·5 was credited. If the donor donated at the kth previous occasion, another $1/k$ was credited. This 'credit' was compounded to a 'score'

applying

$$\text{score} = \sqrt{1 + \sum_{k=1}^{n} I\frac{1}{k}} \qquad (1)$$

$$I = \begin{cases} 1, & \text{if } k\text{th previous donation = yes} \\ 0, & \text{if } k\text{th previous donation = no} \end{cases}$$

with n representing the total number of possible donation attempts at the donor's location. Hence, for all first-time donors the score was equal to 1, since no previous donation attempts were made. The square root transformation was done to achieve a good fit of the logistic link function in the regression analysis. The score gave more weight to recent donations and less weight to older donations.

For example, a first-time-donor has a score of $1 = \sqrt{1 + 0 + 0 + \ldots}$; for a donor who donated five times at all five most recent occasions available to her, the score is $1{\cdot}51 = \sqrt{1 + 1/2 + 1/3 + 1/4 + 1/5 + 0 + \ldots}$; and for a donor with two donations at the current and at the 5th most recent occasion available, the score is $1{\cdot}10 = \sqrt{1 + 0/2 + 0/3 + 0/4 + 1/5 + 0 + \ldots}$

The donor scores for index donations between 1 October 1996 and 30 September 1997 ranged from 1 to 2·327 (mean ± SD: 1·424 ± 0·313; median 1·404, 25%-quartile: 1·157, and 75%-quartile: 1·651). The maximum number of donations among our donors was 150; the 60 most recent donations were compounded in the score by applying eqn 1.

Logistic regression model. In a logistic regression model, the probability of donation, p(D), for a donor was modelled applying two explanatory variables, a score (see eqn 1) for repeat donors and a dichotomous variable, donor_{FT}, indicating a first-time donor status. This model was formally represented by:

$$p(\text{D}) = \text{Prob}(\text{donation}|\text{score, donor}_{\text{FT}}). \qquad (2)$$

The probability of donation and the explanatory variables were linked by the logit function that was selected to model the donor's behaviour and predict her or his p(D):

$$\log(p(\text{D})/(1-p)(\text{D})) = \text{int} + \text{scf}\cdot\text{score} + \text{ftd}\cdot\text{donor}_{\text{FT}} \qquad (3)$$

whence int (intercept), scf (score factor) and ftd (first-time donor) denoted the regression coefficients. The three regression coefficients, int, scf and ftd, can be determined by modelling with actual donation data sets (see next section). By transforming this regression equation, the direct calculation of the probability of donation, p(D), became possible. For ease of use, the equations were shown separately for repeat donors (eqn 4) and first-time donors (eqn 5), but may be compounded in one formula (eqn 6) by using the dichotomous variable donor_{FT}.

The following logistic regression model:

$$p(\text{D}_{ts-te}) = \frac{e^{\text{int} + \text{scf}\cdot\text{score}}}{1 + e^{\text{int} + \text{scf}\cdot\text{score}}} \qquad (4)$$

was selected for repeat donors. t_s denoted time of start and t_e time of end of the donation interval, D.

The value of the score obtained in eqn (1) can be used for calculating the p(D) for repeat donors by applying eqn (4) with the appropriate coefficients for the desired prediction interval (see next section).

First-time donors were dealt with separately to improve the fit of the logistic regression model. The following logistic regression model:

$$p(\text{D}_{ts-te}) = \frac{e^{\text{int} + \text{scf} + \text{ftd}}}{1 + e^{\text{int} + \text{scf} + \text{ftd}}} \qquad (5)$$

was selected for first-time donors.

Of course, eqns 4 and 5 can be combined, if preferred:

$$p(\text{D}_{ts-te}) = \frac{e^{\text{int} + \text{scf}\cdot\text{score} + \text{ftd}\cdot\text{donor}_{\text{FT}}}}{1 + e^{\text{int} + \text{sct}\cdot\text{score} + \text{ftd}\cdot\text{donor}_{\text{FT}}}}$$

$$\text{donor}_{\text{FT}} = \begin{cases} 1, & \text{if first-time-donor} = \text{yes} \\ 0, & \text{if first-time-donor} = \text{no.} \end{cases} \qquad (6)$$

It should be noted that the probability of donating within a certain time frame after an index donation is practically equal to the probability of retesting, because retesting among our donors occurred almost exclusively during the subsequent donation attempts.

Modelling of the logistic regression. With 69 975 donors of the first set, the modelling of a logistic regression was performed to predict the likelihood of a subsequent donation within 170–275 days: $p(\text{D}_{170-275\,\text{days}})$. For a $p(\text{D}_{170-275\,\text{days}})$, the regression coefficients were estimated for first-time donors (ftd = 0·5718) and repeat donors (int = − 3·8410 and scf = 2·4334). The R^2 of Nagelkerke (1991) representing a parameter for the fit of our model was 0·13 and the area under the ROC curve (AUC) representing a parameter for the predictive value (Hosmer & Lemeshow, 1989) was 0·68.

In an ROC curve analysis (Fig. 1) the sensitivity (true-positive rate) is plotted against 1 − specificity (false-positive rate). The area under the ROC curve (AUC) may vary between 0·5, which indicates chance level of prediction, and 1·0, which indicates completely accurate prediction.

Other donor demographic factors, like sex, age, blood group and size of donor community, had some predictive value when used alone (see Table 2). However, they

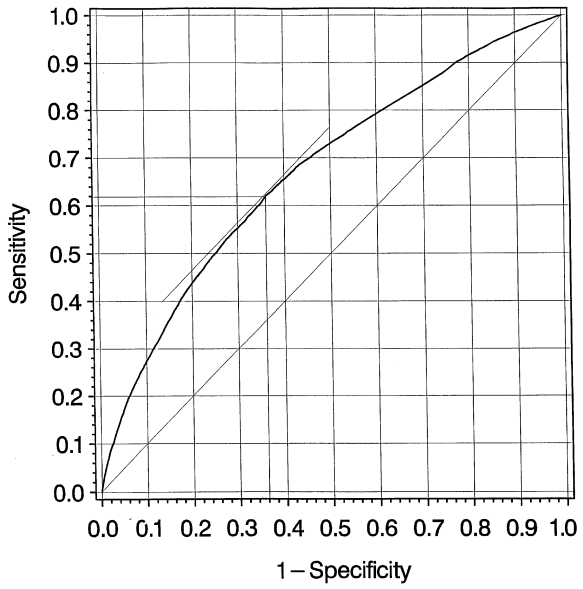

Fig. 1. ROC curve of the logistic regression model for $p(D_{170-275\,days})$. The curve represents unbiased estimates for the sensitivity (true-positive rate) and 1 − specificity (false-positive rate) of the logistic regression model predicting the probability of donation, $p(D_{ts-te})$, during a given time frame (t_s–t_e) with t_s=170 days and t_e = 275 days. For example, the model can attain about 62% sensitivity with 64% specificity. The calculated area under the curve (AUC) is about 0·66. The curve is constructed with the data of an independent, second set of 69 948 donors by using the estimated regression coefficients that were derived from the data of the first set of 69 975 donors. The ROC curves obtained with the regression coefficients of Table 3 were almost identical (not shown).

were not used in the final model because there was practically no improvement of the model's predictive value, when these other demographic factors were used in combination with the donor score (data not shown). The logistic regression model thus obtained was evaluated with 69 948 donors of the second set. As the results of the modelling were considered sufficient (see Results), we did not need to utilize the remaining, third set of donor data for evaluation purposes. It should be noted that there was at least one opportunity for the return visit for each index donation within the desired interval.

Technical procedure for quarantine storage

Blood components, like plasma units, that require quarantine storage were sorted according to their donors' $p(D_{ts-te})$ for one or more preselected time intervals, $t_s - t_e$. According to the known demand for quarantined units, a suitable cut-off value for $p(D_{ts-te})$ was determined for which the expected quarantine unit release equalled or exceeded the demand. We studied the outcome, if blood components of donors with $p(D_{ts-te})$ values equalling or exceeding the cut-off were put into quarantine storage to increase the success rate over the current procedure.

As an example, often the number of plasma units that can be stored will be determined by the given storage capacity (size of −40 °C room). In this case, a $p(D_{cut\text{-}off})$ can be determined for which the number of donations with $p(D)$ values higher than the cut-off equals the number of units that can be stored for the average quarantine period. Only those donations with $p(D)$ values higher than $p(D_{cut\text{-}off})$ will be deposited in quarantine storage. In such a way, the given storage capacity can be used in an optimal way.

If a blood service is planning to establish storage capacity, the required size of a −40 °C room will be determined by the number of plasma units that need to be released after the quarantine storage to meet the demand for patient care. The relationship between units stored and units released (see Fig. 3) is most useful for this purpose. The relationship with consideration given to $p(D)$ values for quarantine storage (Fig. 3, bold line) compares favourably with the standard approach neglecting $p(D)$ values (Fig. 3, fine line).

Hardware and software

All computations were done with the SAS package, release 6·12 (SAS Institute, Carey, NC) for DEC OSF1 on an Alpha Server 1000 A 5/333.

RESULTS

An electronic database comprising 714 182 donors with 4400 965 donation attempts was utilized in this study. The full dataset was evaluated by descriptive and univariate methods (see Tables 1 and 2). Following this description we defined a donor score and developed in the main part of the study a statistical model for the likelihood of donor return based on the donor score. The statistical modelling and its evaluation was done with the subset of donations from October 1995 to September 1996. Then the utility of the statistical model was shown for its effect on quarantine release by applying the model in a retrospective fashion to an independent dataset comprising the donation data available for 1997.

Changes in demographic factors of blood donors

Various donor demographic factors were obtained for all donors and donation attempts in our database and analysed for three equal periods spanning 14 years

Table 1. Donors and changes in their demographic factors since 1985

	May 1985–Oct 1989		Nov 1989–Apr 1994		May 1994–Nov 1998	
Parameter and characteristic	one-time donor*	repeat donor*	one-time donor	repeat donor	one-time donor	repeat donor
Age (years)						
mean ± SD	32·0 ± 12·2	36·8 ± 12·4	33·5 ± 12·2	37·1 ± 12·4	34·7 ± 12·2	37·9 ± 12·2
median (25%-, 75%-quartile)	27·6 (21·6, 41·5)	36·7 (25·1, 47·0)	30·0 (23·4, 42·3)	36·2 (25·9, 48·0)	32·1 (24·6, 43·2)	36·8 (27·7, 47·7)
Sex						
female	41·5%	36·2%	45·8%	39·6%	48·3%	42·4%
male	58·5%	63·8%	54·2%	60·4%	51·7%	57·6%
Residence†						
major cities	6·3%	4·3%	5·3%	3·8%	5·4%	3·4%
towns	24·5%	19·9%	21·1%	19·5%	21·3%	19·9%
villages & countryside	69·2%	75·8%	73·6%	76·7%	73·2%	76·7%
Total‡						
number (*n*)	119 172	258 646	120 293	291 886	124 885	301 188
percentage of total number	31·5%	68·5%	29·2%	70·8%	29·3%	70·7%
fraction of population (15–65 years)§	1·77%	3·85%	1·71%	4·15%	1·77%	4·26%

*Within the indicated time interval. Many but not all of the one-time-donors were first-time donors. †According to 1990 postal code. Major cities: >100 000 inhabitants (Stuttgart, Freiburg, Heidelberg, Heilbronn, Karlsruhe, Konstanz, Mannheim, Tübingen, Ulm); towns (current or former '*Kreisstädte*'): 10 000–100 000; villages & countryside: <10 000. ‡A repeat donor may be counted as one-time donor, if only one donation occurred within a given time period. Many donors were counted repeatedly in two or three time intervals, if they were repeat donors. §In 1989 the population 15–65 years old in the *Land* (state of) Baden-Württemberg was 6718 260 (1993: 7036 736, 1997: 7066 940; Statistisches Landesamt Baden-Württemberg, Stuttgart).

Table 2. Donors in mid 1990s and exemplary demographic factors

Parameter and characteristic	No. of donors who donated within 170–275 days and their proportion among all donors with the same characteristic		
	donors (*n*)*	proportion (%)†	R^2‡
Donor status			0·013
first-time-donor	9277	30·3%	
repeat donor	79 495	44·1%	
Age			0·008
Distribution among females			
18–19 years	865	39·0%	
20–31 years	9345	34·2%	
32–41 years	9352	40·5%	
42–51 years	7409	41·4%	
52–63 years	7026	42·8%	
64–65 years	88	27·6%	
Distribution among males			
18–19 years	807	39·8%	
20–31 years	12 865	38·3%	
32–41 years	14 589	44·7%	
42–51 years	12 500	47·3%	
52–63 years	13 737	48·6%	
64–65 years	189	27·0%	
Sex			0·004
female	34 085	39·0%	
male	54 687	44·3%	
Residence			0·0007
major cities	2173	36·8%	
towns	17 649	40·7%	
villages & countryside	68 950	42·7%	
Blood groups			0·00008**
O Rh neg.	7825	41·5%	
A Rh neg.	8039	42·0%	
O Rh pos.	28 944	42·3%	
A Rh pos.	30 237	42·2%	
AB Rh pos./Rh neg.	4313	42·8%	
B Rh pos./Rh neg.	9413	41·8%	
Total	88 772	42·1%	

*A total of 210 786 donors, who were offered a blood drive and including the 64- and 65-year-old donors, were analysed for being retested within 170–275 days after their first donation between 1 October 1995 and 30 September 1996. For the donor-related statistics presented in this table, only the first donation of donors during this time period was analysed. †Proportion of 'successful' donors among all donors with the specified characteristic. ‡R^2, the coefficient of determination (Nagelkerke, 1991), may be interpreted as the proportion of the explained 'variation' of the outcome. For example, an $R^2 = 0{\cdot}013$ implied that 1·3% of the 'donor's motivation' to attend retesting can be explained by the parameter 'donor status'. **For brevity, the data for Rh pos./Rh neg. were combined in the blood groups AB and B. However, R^2 was calculated for the eight different groups rather than the six groups shown.

(Table 1). An increase of the mean age occurred for repeat donors by about 1 year and was even more apparent for one-time donors. The fraction of women among one-time donors increased to almost 50% in the most recent period. It is worth noting that we were successfully recruiting larger fractions of the eligible resident population as active donors. There may be a trend in the distribution of donors among villages, towns and major cities, as the large fraction of village dwellers among one-time donation attempts has even increased compared with the 1980s.

Possible demographic factors of blood donors affecting successful retesting

We analysed the effect of donor demographic factors on the rate of attending a subsequent donation within a preselected time interval (Table 2). The donor status was the single most informative parameter predicting successful retesting. Age was the second most useful predictor. First-time donors age 18 years were successful in returning, whereas, in particular, female donors in their 20s, possibly as a result of pregnancies and nursing, were less reliable in this respect. The donor's reliability equalled that of 18-year-old first-time donor at about age 30 years and was further and steadily increasing with age until the early 60s.

Male donors returned more often than female (Table 2). Donors from major cities and towns were found to be less reliable than those from villages. We had previously demonstrated by statistical analysis an excess of Rh-negative donors, in particular among female donors (data not shown), at our service (Wagner *et al*., 1995, table 1). We now realized though that Rh-negative donors were somewhat less successful in retesting, probably owing to the observed lower return rate of female donors in general (Table 2). Likewise, the proportion of donations with blood group AB was about average, despite the high demand for plasma units of this blood group.

Besides the difference between first-time and repeat donors, the effects were rather limited and pertained usually to smaller fractions of donors, thus having only limited use for predicting the average likelihood of donation. We thus turned our attention to the donation history, which is very well known for all donors, and considered the possibility of utilizing this parameter for prediction.

Donor score

We postulated that the known donation record of a given donor may indicate her or his likelihood of attempting another donation within a preselected time interval. Using this assumption, we defined a donor score that can be determined for any donor of a blood donation service and depended on the donor's donation record only. This donor score was unique for any given index donation of a donor and would usually be calculated for the donor's latest donation.

Logistic regression model

We developed a logistic regression model to calculate a probability of donation, $p(D_{ts-te})$, within a given time frame (t_s–t_e) as described in the Methods, eqns 4 and 5. The predictive value of our model was confirmed with independent data sets comprising about 69 000 donors (Fig. 1). The donor scores mentioned in the previous section correlated with a $p(D_{170-275\,days})$ of 22% (for a score = 1) and 86% (for a score = 2·327) applicable to any index donation during the 1-year period in 1996/97. All first-time donors had a $p(D_{170-275\,days})$ of 33%. Other demographic factors besides donor score and first-time donor status were not included in the logistic regression model (prediction model) because they did not improve the prediction of the model to any relevant extent (data not shown).

Validation of the logistic regression model

The logistic regression model thus obtained was evaluated with 69 948 donors of the second set. For this independent second set and a $p(D_{170-275\,days})$, very similar coefficients were obtained: ftd = 0·5470, int = − 3·7305, scf = 2·3456. We confirmed the predictive value of our model with this independent set of donors because the values for $R^2 = 0{\cdot}12$ and AUC = 0·67 were almost identical to those values for the first set of donors. An ROC curve for the second donor set is shown in Fig. 1.

Varying index donation intervals and time frames

We applied the model to two different index donation intervals and two different time frames. The coefficients of the model, R^2, which indicates the fit of the model (Nagelkerke, 1991), and the area under the curve (AUC), which represents the predictive value (Hosmer & Lemeshow, 1989) did not change to any great extent (Table 3). This observation gave further evidence for the robustness of our logistic regression model. Hence, we could reliably calculate the $p(D_{ts-te})$ for any donation at our blood service that occurred at least since 1995 and found that the donation history was most powerful in predicting the $p(D)$. The utility of $p(D)$ calculation and the possible

Table 3. Coefficients and statistical parameters of the logistic regression model for two different index donation intervals and two different time frames

	$p(D_{170-275\ days})$				$p(D_{82-365\ days})$			
Index donation interval and variable	first-time donor (ftd)*	intercept (int)*	score factor (scf)*	general description	first-time donor (ftd)	intercept (int)	score factor (scf)	general description
between 1 Oct 1995 and 30 Sept 1996								
parameter estimate†	0·5629	−3·7988	2·4024		0·5253	−3·4287	2·8871	
standard error	0·0163	0·0285	0·0189		0·0146	0·0297	0·0209	
χ^2	1198	17 725	16 183		1294	13 324	19 162	
P-value	0·0001	0·0001	0·0001		0·0001	0·0001	0·0001	
odds ratio	1·76	NA‡	11·0		1·69	NA	17·9	
95% confidence interval	1·70–1·81	NA	10·6–11·5		1·64–1·74	NA	17·2–18·7	
R^2; AUC §				0·12; 0·67				0·15; 0·70
eligible donors (*n*)**				209 768				233 634
proportion of successful donors				42·19%				64·91%
between 1 Oct 1996 and 30 Sept 1997								
parameter estimate	0·5629	−3·5762	2·3114		0·5609	−3·4466	2·9798	
standard error	0·0164	0·0290	0·0191		0·0150	0·0313	0·0220	
χ^2	1173	15 199	14 630		1394	12 109	18 413	
P-value	0·0001	0·0001	0·0001		0·0001	0·0001	0·0001	
odds ratio	1·76	NA	10·1		1·75	NA	19·9	
95% confidence interval	1·70–1·81	NA	9·7–10·5		1·70–1·81	NA	18·9–20·6	
R^2; AUC				0·12; 0·67				0·15; 0·70
eligible donors (*n*)				201 249				228 336
proportion of successful donors				45·04%				67·84%

*Degrees of freedom = 1. †Estimates for the logistic regression coefficients. ‡NA—not applicable. §R^2 represents a parameter for the fit of our model (Nagelkerke, 1991) and AUC represents a parameter for the predictive value of our model (Hosmer & Lemeshow, 1989). **Only those donors were evaluated who had been offered a blood drive at their preferred location during the indicated time frame, excluding all donors older than 63 years.

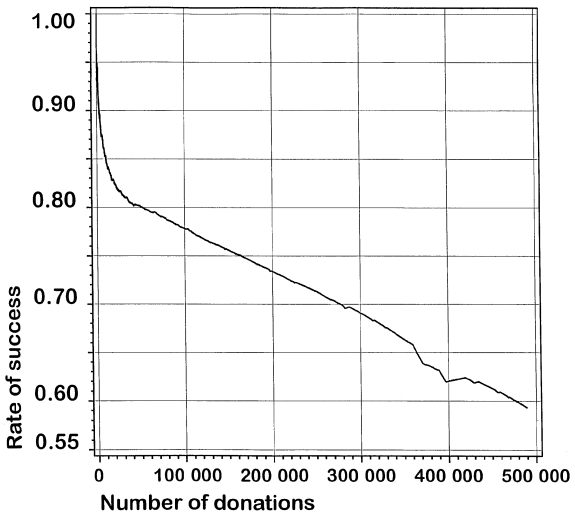

Fig. 2. Rate of quarantine success among all blood donations. At our service, 489 153 units of whole blood were donated in 1997. The donations were sorted according to their calculated $p(D_{170-275\,days})$ and the cumulative number of donations is given on the abscissa. This number increases, when the $p(D_{170-275\,days})$ is decreasing. On the ordinate, the actual rate of success is shown. The donors returned within 170–275 days for 59·3% of all index donations. This percentage represents the average success rate for blood components quarantined at our service in 1997.

application of p(D) for plasma quarantine management was addressed next.

Utility of the calculated $p(D_{170-275\,days})$ to predict the observed success rate

To demonstrate the utility of the calculated $p(D_{170-275\,days})$ for predicting the actual rate of returning, we sorted all donations in 1997 according to their calculated p(D) and plotted the cumulative number of donations against the actual rate of returning within the time interval of 170–275 days (Fig. 2). We obtained a smooth, almost monotonous curve demonstrating a good fit of our model and an excellent prediction. Some very small subgroups were not optimally modelled and caused an increasing rate of success whilst the p(D) decreased. One major deviation was observed around 400 000 donations and was caused by first-time donors, who contributed 59 388 (12·1%) of all donations.

The overall, cumulative success rate averaged 59·3%, which was much higher than the mean donor return rate of 42·11% reported in Table 2. This virtual discrepancy between the success rate calculated for the cumulated donations and the donor return rate is because reliable donors donate more often than less reliable donors.

Behaviour of first-time and repeat donors

The median $p(D_{170-275\,days})$ of all donors was 47% (range 22–86%, 25%-quartile: 33%; 75%-quartile: 61%). The first-time donors' success rate of 33% was much below that median. Hence, the overall success rate increased by 2·5% to 61·8%, if all first-time donations were excluded from quarantine purposes. However, the first-time donors had a better rate of success than some repeat donors with a $p(D_{170-275\,days})$ as low as 22%. The application of our model would hence be far superior to the mere exclusion of plasma units of first-time donors from quarantine storage.

Extended durations of preselected time intervals

Because at the time of this study plasma units expired after 1 year in storage, the time interval of $t_s = 170$ days to $t_e = 275$ days was applied to develop our model. We

Table 4. The effect of varying the length of the time interval (t_s–t_e) with $t_s = 170$ days kept constant*

		Estimates for the logistic regression coefficients				
end of time interval (t_e)	successful donors (%)	first time donor (ftd)	intercept (int)	score factor (scf)	R^2†	AUC†
275 days	42·3	0·54	−3·88	2·49	0·12	0·67
1 years	56·2	0·44	−3·48	2·66	0·13	0·68
1·5 years	71·6	0·33	−3·15	2·99	0·14	0·70
2 years	77·7	0·29	−3·11	3·24	0·15	0·71
2·5 years	81·1	0·19	−2·87	3·25	0·14	0·72
3 years	82·9	0·14	−2·77	3·29	0·14	0·72
3·5 years	84·2	0·08	−2·61	3·26	0·14	0·72
4 years	85·1	0·05	−2·50	3·24	0·14	0·72

*187 435 donors were evaluated with at least one donation between 1 Oct. 1993 and 30 Sep. 1994. †For R^2 and AUC see legend to Table 3.

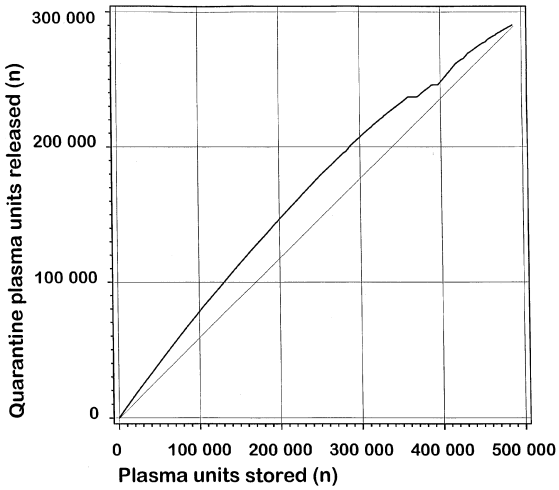

Fig. 3. The relationship of plasma units put in quarantine storage and quarantine plasma units released. The figure relates the number of stored units with the maximum number of units released after quarantine, if storage was done according to the $p(D_{170-275\,days})$ in 1997. The average success rate for all 489 153 donations was 59·3%. The success rate of our quarantine procedure (indicated by the bold line) is considerably higher than this average (indicated by the fine line) for a wide range of quarantine plasma units released. Please note the different scales for the *x*- and the *y*-axes.

were reluctant to shorten t_s because the safety margin for the window period should not be jeopardized. As plasma storage for 2 years and longer may become standard, we tested the effect of extending the time intervals by varying t_e (Table 4). Of course, more donors were successfully donating within the extended time frames and an extension of t_e to 1·5 years would much improve the return rate. However, extending t_e beyond 2 years had only a marginal effect. The relative contribution of first-time donors is much diminished beyond an interval of 2 years. The regression coefficients, first-time donor (ftd), intercept (int) and score factor (scf), can be used to estimate p(D) for a given time interval t_s–t_e or to interpolate the coefficients for varying t_e. Our model was permissive for varying the time frames, because R^2 and AUC increased somewhat, indicating an even better prediction for longer time intervals.

Application of the model to plasma unit quarantine

If the plasma component of all donations in 1997 had stored, 288 600 units representing about 59% of all donations could have been released for transfusion in compliance with the minimum quarantine period of 170 days (Fig. 3). The critical value of p(D) useful as a cut-off for plasma unit storage depends on the total number of stored units. As less than half of all donations were actually stored for quarantine purposes, the success rate of our quarantine procedure could be considerably enhanced above the current 59% average by selecting donations with a high p(D) value attached.

For the supply of the *Land* (state of) Baden-Württemberg with 10 million inhabitants, our blood service produced about 100 000 quarantine plasma units per year. With the current success rate of about 59%, 175 000 units were needed to be put in storage. If the decision for storing any given plasma unit had been based on its donor's p(D) value, we could have retrieved about 130 000 units without incurring any additional costs (Fig. 3). Vice versa, to achieve the release of 100 000 units we would have needed to keep 130 000 units in storage rather than the 175 000 units actually stored per year (Fig. 3). If the storage costs were directly proportional to the number of units stored, the application of our model would have saved about 20% of the storage expenses.

DISCUSSION

Donors give their blood without remuneration, and it is the finest responsibility of the physicians in transfusion medicine to put this gift of life to its optimal use in the interest of the donors and for the benefit of the patients. Driven by this motivation, we developed a procedure to predict a donor's likelihood of donating within a preselected time interval. We showed that the application of our technical procedure could increase the available quarantine plasma by about 30 000 units per year without incurring additional costs or reduce the quarantine storage expenses by about 20% without limiting the current supply. These figures, if widely applicable, may result in a welcome relief to the stringent cost-containment efforts of the health care systems in Germany and elsewhere. We think that the whole demand of the German health care system for single donor plasma units could be met by quarantine plasma and devised means to improve their cost-efficiency.

Previously, donor demographic factors such as age, age at first donation, sex, location of the blood drive and the donation history were reported as being relevant for predicting subsequent donation behaviour (James & Matthews, 1993, 1996). A surprising finding, however, was the consistency of return across age groups, and therefore the limited impact of what has until now been considered a key predictive variable. We have formally shown that the donation history was most useful for predicting the probability of donation, p(D). Because no attempt was made to predict the exact number of days elapsing until the next donation, we did not need to

rely on applying the Cox regression. We could rather adopt the approach of a logistic regression model with a dichotomous variable, successful attending or not attending an offered blood drive at a particular location. Our model may be applicable to other services with similar donor characteristics at temporary donor sites. It would also be possible to determine the donor score for blood centres with the continuous option for donations rather than with blood drives at temporary locations.

As first-time donors were rightfully perceived less likely to return than the average repeat donor (Burnett & Leigh, 1986), their units were often not stored for quarantine purposes. The analysis showed that this decision could improve the overall success rate by about 2·5%. Some repeat donors who are often found among those with small numbers of previous donations were even less likely to return than first-time donors. A rational decision tree for plasma quarantine would relegate donations of this subset of repeat donors first, followed by all donations of first-time donors. Then, donations of another subset of the remaining repeat donors whose *p*(D) did not exceed a certain threshold could be excluded from storage. The required *p*(D) thresholds can easily be determined with considerable precision by our model.

The observed demographic data supported an increase in mean donor age, which has been noted by our profession with disturbance. However, at our service this trend may be virtual and caused in part by the broader participation in our blood drives, because the mean donor age was expected to approach the mean age of the population. This idea was also supported by the observation that the 25%-quartile of donor age increased more than the 75%-quartile; the latter, in particular, did not change much for repeat donors. Women represent now about 50% of one-time donors and their fraction among repeat donors has also increased. This trend could continue, although the childbirth period and menstrual iron loss may prevent an equal contribution by women to the blood supply, who are a very safe cohort among all donors (Piliavin, 1977).

We found trends in certain demographic donor factors that we did not utilize for *p*(D) calculation. Any of these trends in subgroups with distinct donation behaviour might help us to refine our current model. For example, the likelihood of first-time donors to become repeat donors was related to the age at first attempt; a notable exception were first-time donors of 18 years, many of whom appeared to have 'waited' for being eligible and were prone to develop to unusually reliable repeat donors (data not shown). Likewise, the important effect of short-term, temporary deferral has recently been shown (Halperin *et al*., 1998; Piliavin, 1987) and might be introduced as a parameter into our model. However, it may be anticipated that much of these parameters' predictive value was already represented by the score. If these parameters were compounded into our calculations, the *p*(D) may be improved by a fraction of these parameters' isolated effects. The calculated score may be less than optimal for an individual donor, if certain events occurred, like change of address or pregnancies; in the aggregate analysis (Figs 2 and 3), however, less than optimal scores were compounded with 'better' scores and the application to the actual datasets confirmed the suitability of 'donor score' even if some scores were – and ever will be – predicted in a less than optimal way.

As more plasma units of blood group AB are needed compared to those of other blood groups, different thresholds may be defined reflecting the relative demand. Several of the rare red cell units, e.g. O CCDee KK, are in short supply but still plentiful enough that not nearly all units need to be stored frozen in liquid nitrogen. The *p*(D) may be used to define cut-off values for the management of blood donations from such 'rare' donors.

The *p*(D) lends itself to reconstruction of the aggregate behaviour of donors at particular times or locations, which was recently shown to be rather predictable (Whyte 1999). The 'bottom-up method' by summarizing the individual *p*(D) values for predicting the aggregate behaviour at particular times or donation localities may be more useful than the common method of observing the aggregate behaviour only (Whyte, 1999). For example, the donation opportunities may be enhanced at those locations where the donor population was shown to be more eager to participate by the aggregate *p*(D) donor values of particular donation sites (Burnett & Leigh, 1986). At a minimum, our method would add a second line of evidence to the established procedures for the planning of blood drives and of an adequate blood supply.

For the current demand of about 100 000 quarantined plasma units per year, we devised a method to improve the rate of recovery to about 78% from the current average of 59%. Still, a 78% success rate may be considered far from optimum. Our observations indicated that there was much room left for studying and improving our understanding of the donor behaviour. However, the R^2 values in Table 3 were about 0·12, which implied that about 12% of the 'donor's motivation' to donate within the preselected time interval can be explained by our model using the combination of the two-parameter 'donor score' and 'donor status'. This figure compared very favourably with the R^2 in Table 2 showing that the parameters age, sex, residence and blood groups covered less than 1% each. In addition to these latter parameters of minor contribution, other,

more relevant, factors may comprise the attending of blood drives in groups, such as social clubs or families (Mayo, 1992; Parker, 1996); these data may be hidden in our data files and await for 'data mining'. Further factors, like multiple donors among the peer group and the family members or psychological attitude favouring donation (Callero & Piliavin, 1983) that were not easy to quantify and may not be practical to determine as routine procedure, may remain beyond our resources for blood supply management for some time to come.

ACKNOWLEDGMENTS

We wish to thank Hans Stegmann for donor database queries, Dr Rainer Muche for helpful statistical advice and reading the manuscript, and Dr Markus Wiesneth for reading the manuscript. We acknowledge the secretarial assistance of Mrs Vanessa Ries.

REFERENCES

Burnett, J.J. & Leigh, J.H. (1986) Distinguishing characteristics of blood donor segments defined in terms of donation frequency. *Journal of Health Care Mark*, **6,** 38–48.

Callero, P. & Piliavin, J.A. (1983) Developing a commitment to blood donation: the impact of one's first experience. *Journal of Applied Social Psychology*, **13,** 1–16.

Cox, D.R. & Oakes, D. (1984) *Analysis of Survival Data*. Chapman & Hall, London.

Halperin, D., Baetens, J. & Newman, B. (1998) The effect of short-term, temporary deferral on future blood donation. *Transfusion*, **38,** 181–183.

Hosmer, D.W. & Lemeshow, S. (1989) *Applied Logistic Regression*. John Wiley & Sons, New York.

James, R.C. & Matthews, D.E. (1993) The donation cycle: a framework for the measurement and analysis of blood donor return behaviour. *Vox Sanguinis*, **64,** 37–42.

James, R.C. & Matthews, D.E. (1996) Analysis of blood donor return behaviour using survival regression methods. *Transfusion Medicine*, **6,** 21–30.

Mayo, D.J. (1992) Evaluating donor recruitment strategies. *Transfusion*, **32,** 797–799.

Nagelkerke, N.J.D. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78,** 691–692.

Parker, M.E. (1996) Multi-gallon blood donors [letter]. *Transfusion*, **36,** 288.

Piliavin, J.A. (1990) Why do they give the gift of life? A review of research on blood donors since 1977. *Transfusion*, **30,** 444–459.

Piliavin, J.A. (1987) Temporary deferral and donor return. *Transfusion*, **27,** 199–200.

Wagner, F.F., Kasulke, D., Kerowgan, M. & Flegel, W.A. (1995) Frequencies of the blood groups ABO, Rhesus, D category VI, Kell, and of clinically relevant high-frequency antigens in South-Western Germany. *Infusionsther Transfusionsmed*, **22,** 285–290.

Whyte, G. (1999) Quantitating donor behaviour to model the effect of changes in donor management on sufficiency in the blood service. *Vox Sanguinis*, **76,** 209–215.