Big data modeling to predict platelet usage and minimize wastage in a tertiary care system

Leying Guan1,a, Xiaoying Tian1,a, Saurabh Gombarb, Allison J. Zemekb, Gomathi Krishnanc, Robert Scottd, Balasubramanian Narasimhana, Robert J. Tibshirani1,b,2, and Tho D. Phamb,2,3,4

*Department of Statistics, Stanford University, Stanford, CA 94305; Department of Pathology, Stanford University, Stanford, CA 94305; 2Stanford Center for Clinical Informatics, Stanford University, Stanford, CA 94305; 3Stanford Hospital Transfusion Service, Stanford Medicine, Stanford, CA 94305; 4Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and 5Stanford Blood Center, Stanford Medicine, Stanford, CA 94305

Contributed by Robert J. Tibshirani, August 10, 2017 (sent for review June 25, 2017; reviewed by James Burner, Pearl Toy, and Minh-Ha Tran)

Maintaining a robust blood product supply is an essential requirement to guarantee optimal patient care in modern health care systems. However, daily blood product use is difficult to anticipate. Platelet products are the most variable in daily usage, have short shelf lives, and are also the most expensive to produce, test, and store. Due to the combination of absolute need, uncertain daily demand, and short shelf life, platelet products are frequently wasted due to expiration. Our aim is to build and validate a statistical model to forecast future platelet demand and thereby reduce wastage. We have investigated platelet usage patterns at our institution, and specifically interrogated the relationship between platelet usage and aggregated hospital-wide patient data over a recent consecutive 25-mo period. Using a convex statistical formulation, we have found that platelet usage is highly dependent on weekday/weekend pattern, number of patients with various abnormal complete blood count measurements, and location-specific hospital census data. We incorporated these relationships in a mathematical model to guide collection and ordering strategy. This model minimizes waste due to expiration while avoiding shortages; the number of remaining platelet units at the end of any day stays above 10 in our model during the same period. Compared with historical expiration rates during the same period, our model reduces the expiration rate from 10.5 to 3.2%. Extrapolating our results to the ∼2 million units of platelets transfused annually within the United States, if implemented successfully, our model can potentially save ∼80 million dollars in health care costs.

prediction | blood | supervised learning

Blood products, including red blood cells, plasma, and platelets, are therapeutics essential to daily in-patient hospital care. Many modern medical interventions would be rendered impossible without these precious resources. Although it is a certainty that these products are needed in any hospital setting, it is uncertain how much and when each product is needed. Additionally, all these products are perishable, the shortest practical shelf life in many cases being only 3 d for platelets. Platelet transfusions are generally used to mitigate or prevent bleeding in a variety of patient populations, including actively bleeding patients (e.g., trauma, surgery) and patients with dysfunctional or insufficient platelets. This combination of uncertain quantity of demand, product perishability, and expense makes predicting hospital-wide platelet usage a high-priority endeavor for all hospitals and blood banks.

Key goals for blood product management in health care systems are to both (i) always have enough products available to keep up with patient demand even in cases of emergencies and (ii) minimize the number of products wasted due to expiration. The primary aim of medical professionals is always to optimize patient care, leading transfusion services and blood centers to err on the side of overstocking, resulting in the expiration and wastage of some platelet products. Provided by volunteer donors, platelets are processed by community blood centers that manufacture, test, and distribute these blood components to their hospital customers. Platelet products donated by single donors (apheresis platelet) are the most challenging and costly to collect, store, test, and manage, from an inventory perspective (1–3). Difficulties include fewer apheresis platelet donors than whole blood donors, longer donation process than a whole-blood donation, less capacity for collection due to more complicated and costly equipment required, and a practical shelf life of only 3 d, due to 2 d sequestration in testing (1, 4, 5). Furthermore, because products are not available on the day of collection and because daily usage is quite variable, collection strategies must attempt to match current inventory levels with highly variable and unpredictable future usage.

Therefore, a common strategy employed by many blood centers is to aim for an expiration rate of 8 to 9%, with 2 million units transfused annually and the outdated rate for apheresis platelets in 2013 being 11.0% of total units produced (6), effectively producing a surplus buffer so as to always have enough platelet units on the shelf for unanticipated patient demand. However, in large tertiary care settings where close to 15,000 units of platelets are potentially transfused annually, this wastage may translate to close to $1 million in expiration alone, depending on the geographically determined price of a unit of platelets. In addition to the cost of the product, there are also indirect costs such as staffing/nursing costs to receive and return the products, and opportunity costs of donors’ hours of donation time. Although the wastage appears large, it is seen as a necessary precaution to have enough products on hand at all times to meet the highly variable and unpredictable patient demand.

Hospitals and blood banks have not yet developed a feasible method for predicting fluctuations in product usage. We aim to build and validate a reliable statistical model for blood banking

Significance

In modern hospital systems where complicated, severely ill patient populations are the norm, there is currently no reliable way to forecast the use of perishable medical resources to enable a smart and economic way to deliver optimal patient care. We here demonstrate a statistical model using hospital patient data to quantitatively forecast, days in advance, the need for platelet transfusions. This approach can be leveraged to significantly decrease platelet wastage, and, if adopted nationwide, would save approximately 80 million dollars per year. We believe our approach can be generalized to all other aspects of patient care involving timely delivery of perishable medical resources.

Author contributions: S.G., A.I.Z., and T.D.P. designed research; S.G., G.K., R.S., and B.N. performed research; L.G., X.T., R.J.T., and T.D.P. analyzed data; and L.G. and X.T. wrote the paper.

Reviewers: J.B., University of Texas Southwestern; P.T., University of California, San Francisco; and M.H.T., University of California, Irvine.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

1. G.K. and E.A. contributed equally to this work.

2. To whom correspondence may be addressed. Email: tibs@stanford.edu or thopham@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1714097114/-/DCSupplemental.
that would have a highly significant impact in improving quality, safety, access to care, and cost of care. Measures to increase ef-
ficiency and cost-effectiveness in health care are becoming in-
creasingly important (7, 8), and an accurate predictive algorithm
for blood product usage could greatly contribute to value-based
care in transfusion medicine.

Due to the critical need for platelets nationwide (9), and the
particularly short shelf life of platelet products, there is an unmet
need for a system that can forecast platelet usage in a hospital
setting. In the care of each patient, the decision to transfuse
platelets is not random. Previous studies have investigated pa-
tient parameters in specific clinical scenarios as an indicator for
platelet usage (10–15). Many of these have focused on small
patient populations, and manual chart reviews. Among other
results, one study indicated that day-of-the-week status could
help in predicting platelet usage (14).

Here, we forecast hospital-wide platelet needs a priori by
leveraging the “big data” available within modern hospital elec-
tronic medical records (EMR). We hypothesized that looking into
hospital-wide patient data associated with clinical transfusion de-
cisions would allow us to accurately predict platelet usage 3 d in
advance, allowing ample time for blood centers to alter collection
and importing strategies. Our institution, Stanford Blood Center
(SBC), is the community blood center that provides all blood
products used at both our associated hospitals, Stanford Health
Care (SHC) and Lucille Packard Children’s Hospital (LPCH). We
have developed a statistical model at our institution, using shared
clinical data, that forecasts platelet demand 3 d in advance, and
reduces combined wastage from expiration by the blood center
and transfusion service. Using recent historical data comprising
29 consecutive months, our model reduces the expiration rate by
two-thirds (from 10.5 to 3.2%) over the same period. In addition
to reducing expiration due to wastage, this model prohibits platelet
shortages; there are at least 10 units remaining on the shelf at the
end of each day, thereby never compromising patient care. With
an annual transfusion rate of ~13,000 units at our institution, this
could potentially reduce our wastage by 950 units annually without
compromising patient care. Extrapolating to total platelet usage of
2 million units in the United States per year, if implemented suc-
cessfully, this approach could save 80 million dollars in annual
health care costs.

Materials and Methods

Inventory Flow. The majority of blood products collected at SBC is distributed
to the Transfusion Service located in the SHC adult hospital (SHC-TS). SHC-TS
acts as a central transfusion hub from which blood products are distributed to
various units and locations within SHC, as well as to LPCH. All of the data for
transfusible products, to both SHC and LPCH, are managed and recorded by
SHC-TS. Additionally, essentially all of the blood products issued for trans-
fusion through SHC-TS are supplied by SBC.

SBC EMR Data Acquisition. We obtained Stanford IRB approval to perform chart
review (Protocol ID 350350) of SBC EMR. Through the Stanford Translational
Research Integrated Database Environment, we obtained SHC data within the
date range January 1, 2013 through June 1, 2015. The dataset includes complete
blood count (CBC) and hospital census by location. We computationally cleaned
the data and aggregated them into a useable relational database for further
analyses. In combination with day-of-week status and number of transfused
products from the previous week, these variables account for 43 covariates
(Tables 1 and 2 show the details for covariates other than day-of-week sta-
tus). From LPCH, CBC and hospital census data were unavailable for research
purposes during the same time period.

Transfusion and Expiration Data Acquisition. All of the transfusion data and a
portion of the expiration data within the specified date range were obtained from
the SafeTrace X (Haemonetics) database system at SBC-TS. These data rep-
represented transfused products for both SHC and LPCH. For each product,
we obtained the donation identification number, issue date and time, ex-
piration date and time, discard date and time, ABO blood group status, and
Rh status. Due to the inventory management procedure in place between SBC
and SHC-TS, a large fraction of the expired platelet products is shipped back to
SBC. Therefore, total platelet wastage due to expiration was determined by
combining the number platelet units that expired at SBC and SHC-TS. Expired
platelet products at SBC were obtained from the SafeTrace (Haemonetics)
system at SBC.

Statistical Modeling. Our goal is to minimize platelet wastage while ensuring
we have enough platelet units to meet patient demand every day. Since the
objective is not prediction but an optimal ordering strategy, standard sup-
ervised learning is not ideal for this problem. Hence we developed a con-
strained optimization strategy. To achieve our goal, accurate prediction for
the next day is not sufficient, as the acquired platelets spend the first 2 d in
testing and will only be available for transfusion on the third day. To ade-
quately account for the demand in 3 d, we instead built a prediction model
for the platelet usage for the next 3 d, with an objective to minimize the
waste. Moreover, on each day, we aim to collect a number of platelet units
such that (i) we will have sufficient platelets to cover the predicted demand
and (ii) we will have enough supply on the shelf in cases of unpredictable
emergency situations.

To formalize our model, we first introduce the following notation. Let $y_i$
be the number of platelet units used on day $i$, $x_i$ be the number of fresh
platelet units arriving on day $i$, $r(1)$ be the number of platelet units that expired on
day $i+1$, $r(2)$ be the number of platelet units that expired on day
$i+2$, $z_i$ be the vector of 43 covariates measured on day $i$, and $w_i$ be the
number of platelet units wasted on day $i$. For our model, it was important
to forecast the platelet demand at the end of 3 d into the future. Collection occurs
on day $i+1$, and platelets will be available for use on days $i+1$ through $i+3$. Therefore,
on day $i$ (i.e., 1 d before collection and 3 d before product avail-
ability), a forecast would be helpful, since there is still actionable time
to maneuver and adapt certain donor recruitment strategies.

Using the 43 covariates available from the hospital data, we predict the
platelet usage for the next 3 d to be

$$ t_i = x_i^T \beta, $$

where $\beta$ is the vector of coefficients for each of the covariates. We solve the
following optimization problem,

$$ J(\beta) = \sum_{i=1}^{n} w_i^T + \sum_{i=1}^{n} r(1)_i + \sum_{i=1}^{n} r(2)_i $$

subject to

$$ 3 \text{ d's total need} = x_i^T \beta, $$

new arrival (collecting strategy) $x_{i,3} = r(1)_i - r(2)_i - x_{i-1} - x_{i-2}$,

$$ \text{waste} w_i \geq \lvert r(1)_i - y_i \rvert, $$

actual remaining $r(1)_i = (r(1)_{i-1} + r(2)_{i-1} - y_i - w_i)$,

$$ r(2)_i = \lvert x_i - y_i - w_i - r(1)_i - z_i \rvert, $$

fresh units remaining $r(2)_i \geq c_i$. 

Here the subscript $+$ indicates a positive value, such that $|x_i|$ is equal to $x$ if
$x > 0$ and zero otherwise. Eq. 1 is the objective we want to minimize, which is
the number of total wasted units with a sparsity penalty on the underlying
coefficients, similar to the lasso method (16). Eq. 2 is the predicted total units
for the next 3 d using the 43 predictors detailed above. Eq. 3 gives a col-
lection strategy for platelets that ensures we collect enough platelet units to
cover the third day’s predicted usage. Eqs. 4-6 detail the calculations for the
wasted units and leftover units. For Eq. 4, the original mathematical definition
of the waste is an equality, but we have relaxed it to an inequality to make the
problem convex. In Eq. 7, we place a constraint that the remaining freshly
received platelets on the shelf should not be less than a certain threshold $c_i$.
For a major hospital with ~40 daily platelet transfusions, we set the threshold
at 30. The index is shifted 3 d into the future because the units we decide to
collect on day $i$ will not be available for transfusion until day $i+3$.

The above optimization problem can be expressed as a convex optimi-
ization problem with linear constraints. We solve the above problem using
the R language library IpSolve (17). The penalty parameter $z$ is chosen through
eightfold cross-validation.
Results

Platelet Transfusion and Expiration Patterns. Within the specified date range, we obtained more than 120,000 transfusion records. Of these 120,000 transfusions, ∼30,000 were platelet transfusions, with platelet transfusions occurring on every day over the 29-mo period. As shown in Fig. 1A, daily platelet usage was highly variable, showing a transposed daily average of 35.4 units and a standard of deviation of 9.2. To gain a better understanding of the daily usage patterns of platelets, we investigated whether daily usage varied across different months or day of week (Fig. 1B and C). As expected, platelet usage was statistically significantly higher on the weekdays compared with weekends (mean 38.4 vs. 27.8, P value = 2.2e-16). These results are summarized by day of week in Table S3. There was a significant difference in daily usage rates across different months, as determined by one-way ANOVA (F = 4.441, P = 1.57e-06).

We additionally explored the rate of daily platelet wastage due to expiration. Similar to the daily variation in transfused products, we also found that platelet expiration varied daily (Fig. 2A). On average, 3.7 units of platelets were expired per day, with a SD of 6.7. We next interrogated daily expirations by day of week and month (Fig. 2B and C). There were statistically significant differences in wastage when comparing weekday to weekend (weekday mean = 2.56, weekday mean = 4.09, P = 5.4e-04) as well as among the different months as determined by one-way ANOVA (F = 4.9, P = 2.16e-07).

During this 29-mo period, ∼3,200 units expired either at SHC-TS or SBC. This translates into an expiration rate of 10.3% during this 29-mo period, containing both the training and validation sets (discussed in Minimizing Platelet Expiration). For the time period corresponding to the validation set, the expiration rate was 10.5%.

Predictors of Platelet Usage. To accurately predict platelet usage and minimize wastage, we first investigated which covariates within the data had strong correlations with platelet usage. We have CBC data for the SHC patients for every day during the 29-mo period, containing both the training and validation sets (discussed in Minimizing Platelet Expiration). The time period corresponding to the validation set, the expiration rate was 10.5%.

We solved the optimization problem detailed in Statistical Modeling and obtained the following values: (i) β, a vector of length 43 containing the weights for the 43 covariates used in the model, and (ii) x̄, the number of fresh platelet units coming on day i.

The t β penalty used in the minimization (Eq. 1) enforces sparsity in the 43 coefficients. Therefore, only 35 of the 43 weights were estimated as nonzero. In Table S4, we list the top 16 coefficients from the CBC and census data, in addition to day-of-week status and previous week's platelet usage, with the highest absolute weights.

![Fig. 1. Daily platelet usage patterns. (A) Daily usage from January 1, 2013 through May 31, 2015 is highly variable. (B) Mean daily platelet usage is higher on weekdays than weekend (38.4 vs. 27.8, P value = 2.2e-16). (C) Mean daily platelet usage varies with month as determined by one-way ANOVA (F = 4.44, P = 1.57e-06).](image-url)
We see from Table S4 that the most prominent predictors are the average platelet usage for the last week (4.21) and the number of patients in the H1 unit, housing Neurosurgery, Trauma, and Vascular patients. It is also worth noticing that there is a clear weekday/weekend effect. The platelet usage for the next 3 d is substantially lower if it covers the weekend (Friday: −3.14), and starts increasing on weekdays (Sunday: 1.95). Surprisingly, the number of patients with low PLT was assigned a negative weight; we would expect that more patients with abnormally low PLT would require more transfusions. However, as shown in Table S4, one of the strongest predictors of future platelet usage is the platelet usage during the past 7 d. Furthermore, as shown in Fig. 3, there is a correlation between PLT and red blood cell count (RBC). Therefore, the negative weight of low PLT on future platelet demand is the high correlation of the low PLT measurement with the RBC measurement as well as the platelet usage during the past week, both of which have already been assigned positive weights. In fact, when we look at the marginal correlations

![Image](https://example.com/image)
(Fig. 3), we can see that low PLT measurements have a weak positive correlation with platelet usage.

**Minimizing Platelet Expiration.** For subsequent analyses involving platelet usage prediction, we used a dynamic training and prediction model. We first split the data into training and validation sets. For the training set, we used all data from the first consecutive 200 d. Using the base model established from the training set, we proceeded to predict platelet demand and calculate appropriate collection volumes every day from day 201 onward. To account for potential changes in patient populations and practices during the period, we retrained our model every 7 d. For example, we start with the first 200 d to build our model and make predictions for number of units to collect per day from day 201 to day 207. We then retrained our model by incorporating the data from days 201 through 207, and then used this updated model to make daily predictions for number of units to collect from day 208 to day 214. This process continues until day 880, the end of our study period. Therefore, our validation set is considered to be days 201 through 880, since each daily prediction for that timeframe was based on a model that did not previously include that day.

By solving the optimization problem as described in *Statistical Modeling* and following our testing procedure described above, we obtained a collection strategy to collect $x_{i+3}$ units on each day to ensure enough supply for the third day in the future. Adopting this strategy and testing it on the validation set, we find that we can reduce wastage to 3.2%, compared with the historical waste of 10.5% over the same time period as the validation set. Annualized, this corresponds to a reduction in the number of wasted platelet products from $\sim$1,400 per year to just 430 per year.

Most importantly, although we significantly reduced wastage from expiration, we ensured that there is never a day when platelet inventory falls short and that there is always enough supply on the shelf every day. We determined the end-of-day remaining units and the wasted units on our modeled collection strategy on the validation set. Fig. 4 illustrates the actual daily units transfused during the validation period, overlaid with our modeled end-of-day remaining units and wasted units. Using our modeled collection strategy, there was not a single day when we were short of inventory, when patients’ platelet demand outpaced platelet supply. Additionally, at the end of each day, there was always a healthy remainder of platelet units on the shelf, with the lowest being 10 units remaining on hand.

In Fig. 5, we further illustrate the cumulative units wasted over a 2-y period in comparison with the cumulative number of units transfused. There is a substantial reduction in wasted platelet units, using our modeled method compared with the historical wastage during the same time period. We also see that the number of wasted units is less variable over time under our modeled strategy.

**Discussion**

Blood products remain essential to modern medicine, and it is crucial to maintain enough supply within each hospital’s transfusion service to adequately meet patient demand, which has been mostly unpredictable. Due to their short shelf life, platelet products are most severely affected by fluctuations in usage. Combined with a high cost of production, storage, and testing, appropriate inventory management of platelet products is even more crucial in
light of their therapeutic indications and efficacy. We demonstrate a statistical model that can adequately forecast patient platelet demand to guide inventory management and decrease wastage due to expiration, while avoiding platelet shortages.

At our institution, where ∼13,000 apheresis platelet products are transfused annually, the daily usage variation is high, with the daily average being 35.4 units per day, and a 25.9% coefficient of variation. These initial findings demonstrate the difficulty of accurately estimating platelet usage. Approximation cannot be based solely on historical mean daily usage. The current strategy employed at our institution is approximating by daily median usage, resulting in an historical expiration rate of 10.3% during this studied time period. More importantly, historical transfusion rates cannot account for changes in patient population, physician practices, or other factors that can dynamically influence platelet usage.

Based on these considerations, we investigated the relationship between platelet usage and in-patient hospital data, including census, CBC, day of week, and number of platelet units transfused during the previous week. We interrogated a total of 43 covariates and found 35 covariates to be significant for predicting platelet usage, using $r^2$ to enforce sparsity. The strongest predictors of future platelet usage are the number of platelets transfused the previous week, the number of patients within our hospital’s sublocation H1 (which includes the Trauma, Neurosurgery, and Vascular Care patients), and the day of the week.

The data furthermore indicate that this relationship can be leveraged to predict future usage. From this, we developed a model that would guide current platelet collection strategy based on a predicted future platelet demand, thus allowing some time for optimizing donor recruitment strategies. The model prohibits a shortage of platelets for any day while minimizing the number of expired products. We used a dynamic and predictive model that could account for potential changes in patient populations and transfusion practices. Using our model, the expiration rate was reduced to 3.2% compared with the historical 10.5% over the same period. At our institution, this difference would result in ∼950 units saved annually. If successfully implemented and extrapolated nationwide where there are ∼2 million units of platelets transfused, with an approximate 11% expiration rate, our approach can potentially save 80 million dollars in health care costs. Furthermore, and most importantly, this reduction in wastage is accomplished while still maintaining a robust inventory of platelets at the end of inventory for each day, in the event of unpredictable surges in patient demand (e.g., trauma).

Our model can be further refined by obtaining patient data from LPCH (our institution’s pediatric hospital) to accompany the transfusion data. There are technical difficulties in obtaining these data. One of the more obstinate obstacles in acquiring LPCH patient data during the specified date range is that the LPCH EMR changed from one system to another. We are uncertain how this change would affect the quality and consistency of our data. Furthermore, the number of LPCH platelet transfusions account for a minority compared with SHC. From these considerations, we decided to not include a data source that could be unreliable and would only account for a minority of transfusions. Additionally, our model uses aggregated data to minimize the lack of LPCH patient data.

Although our model is currently specific to platelets transfusion and wastage, this strategy can be extended to other blood products without difficulty. We focused on platelets initially because, if operationalized, wastage due to expiration of this product would see the most significant impact. Additionally, with anticipated regulatory guidance on the horizon relating to platelet storage (18–25), there can be additional stress on the platelet inventory of the entire country. With this potential regulatory change being finalized, it becomes even more crucial to manage platelet inventory effectively to minimize wastage and maintain patient care.

Furthermore, the strategy we used can be generalized to be applied to other health care-related products that have the following criteria: (i) always have to be on hand, (ii) have high costs, and (iii) are perishable (e.g., Epinephrine Auto-Injectors). If generally applied, this method represents a logical and automated approach to changing the way health care is delivered on a system-wide level. It can contribute to significant cost savings in the health care setting without ever compromising patient care. We are encouraged by the potential that these findings represent.

The first step to minimizing wastage while maintaining top-of-the-line patient care is to understand the entire supply chain, from donor collection to patient usage. The relationship among SBC, SHC, and LPCH allows for a more fully transparent view of data to enable this type of vein-to-vein analysis. Our findings promise the potential to customize, in real time, donor recruitment strategies based on anticipated future patient demand. We plan to implement this data-driven process at our institution in the near future, using the findings herein as a starting point. We will also distribute an open source software package so that other hospitals and blood centers can implement this process.

ACKNOWLEDGMENTS. We thank Drs. Hua Shan, Scott Boyd, and Laura Lazzeroni for their helpful comments and discussions. R.J.T. was supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Grant N01-HV-28183.