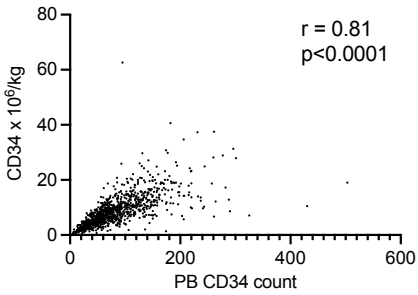
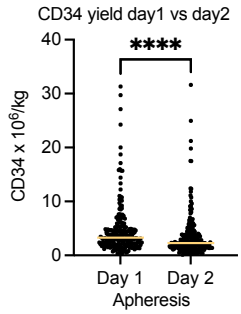
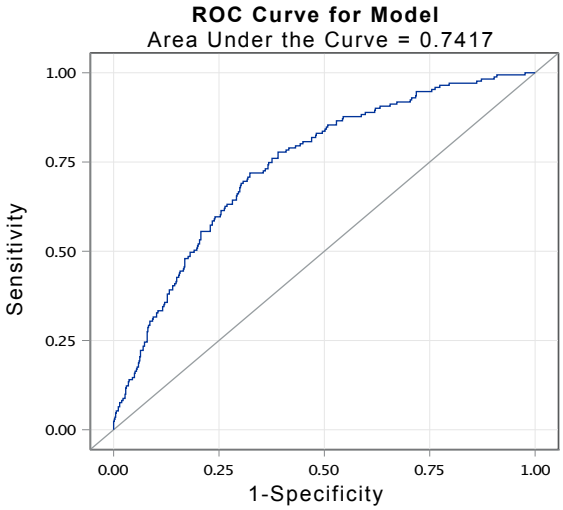
Supplemental Figure 1. Correlation between Day 1 PB CD34 count with Day 1 CD34 yield in G-CSF mobilized donors (N= 1023).



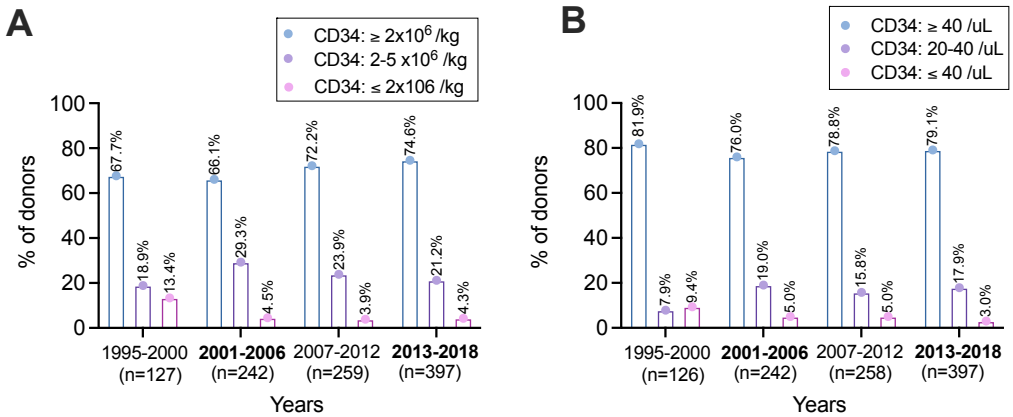
Supplemental Figure 2. Day 1 PB CD34 count on Day 1 vs Day 2 of LP. 213 out of 1025 G-CSF mobilized donors did not achieve adequate stem cell yield goal and collected for an additional day (day 2).



Supplemental Figure 3: ROC curve for logistic model



Supplemental Figure 4: Percentage of "poor", "less-than-optimal" and "good" mobilizer from 1995-2018 in G-CSF mobilized donors.



Supplemental Table 1. Characteristics of the demographics and clinical features in G-CSF mobilized donors (N=799).

Variables/Features	CD34 < 40 /μL	CD34 ≥ 40 /μL
Total N	171	628
WBC	5.9 (2.5-12.5)	6.7 (3.2-15.5)
Age	55 (18-77)	50 (18-76)
Female	92 (53.8%)	270 (43%)
BMI	26.95 (17.8-53.94)	29.72 (16.14-65.71)
Underweight	0 (0%)	3 (0.5%)
Normal	60 (35.1%)	116 (18.5%)
Overweight	65 (38%)	208 (33.1%)
Obese	46 (26.9%)	301 (47.9%)
RBC	4.51 (2.88-6.57)	4.74 (3.53-6.44)
Hemoglobin	14 (9.9-18.8)	14.35 (9.8-17.9)
Hematocrit	41.6 (32.2-56.3)	42.5 (31.3-53.4)
Platelets	226 (113-430)	243.5 (97-644)
MCV	92 (38-111.8)	89.65 (66.4-111)
MCH	31.1 (20.4-103.4)	30.3 (21.1-37.4)
MCHC	33.9 (30.8-35.8)	33.7 (30.7-35.8)
MPV	8 (6.3-11.5)	8.2 (5.9-11.1)
Neutrophil, absolute	3.5 (0.8-9.5)	4 (0.9-11.2)
Lymphs, absolute	1.7 (0.4-3.3)	1.9 (0.7-3.8)
Mono, absolute	0.4 (0.1-0.8)	0.5 (0-1.2)
EOS, absolute	0.1 (0-0.7)	0.1 (0-1.8)
BASO, absolute	0 (0-0.2)	0 (0-0.3)
Neutrophil	60.5 (34-85)	60.6 (23-88.9)
Lymphocyte	29 (7.8-53)	28.7 (8.8-72)
Monocyte	7.1 (2-14)	7.2 (1-15.8)
Eosinophil	2 (0-9)	2.1 (0-27.9)
Basophil	0.7 (0-2.8)	0.7 (0-3.7)
Sodium	141 (133-146)	140 (131-146)
Potassium	4.1 (3.1-5.2)	4 (1-9.9)
Chloride	104 (94-110)	104 (96-112)
CO_2	28 (22-36)	28 (16-36)
AnionGap	9 (2-16)	9 (3-25)
BUN	14 (5-30)	14 (5-51)
Creatinine	0.8 (0.4-1.63)	0.83 (0.4-2.79)
Bilirubin, total	0.4 (0.2-1.3)	0.4 (0.1-1.9)
Uric Acid	4.6 (1.9-11.7)	5.5 (1.4-11.1)
Glucose	101 (58-299)	100 (51-411)
Calcium	9.6 (8.8-10.5)	9.6 (5-11.6)
Protein, total	7.5 (6.5-9.1)	7.5 (6.3-9.2)
Albumin	4.4 (3.9-5.4)	4.5 (3.5-5.4)
Alk Phos	74 (30-146)	71 (20-191)
ALT	25 (9-126)	29 (7-198)
AST	23 (7-78)	24 (8-119)
LDH	170 (103-660)	172 (70-354)
PT	12.4 (9.5-22.9)	11.8 (9.4-46.6)
INR	1.02 (0.85-1.92)	1.02 (0.84-27.5)
PTT	29.9 (20-38.3)	30.1 (20.9-51.6)
Day 1 CD34 count /μL	28 (6-39)	80 (40-264)
Day 1 CD34 /kg of b.w.	3.13 (0.7-8.9)	8.85 (0.77-62.65)

Supplemental Table 2: Donor-recipient match in the subset of the G-CSF mobilized donors (N=799)

Donor-Recipient Match	N	%
Identical siblings	564	70.60%
Haploidentical donors	232	29%
Missing	3	0.40%

Supplemental Table 3: Multivariate Logistic Regression Model

		CD34<40			
Covariate	Mean±SD	Odds Ratio*	95%CI Low	95%CI Up	P-value
Age (years)	**48.8±13.6**	1.44	1.18	1.74	**<.001**
BMI	**30.2±6.6**	0.64	0.51	0.81	**<.001**
Platelets	**246.4±59.6**	0.59	0.47	0.75	**<.001**
MCH	**30.4±2.3**	1.62	1.16	2.26	**0.005**
MPV	**8.2±0.9**	0.75	0.61	0.92	**0.007**
ALT	**33.2±18.1**	0.61	0.47	0.79	**<.001**

Number of observations in the original data set = 799; Number of observations used = 799; * OR representing the effect per 1-SD increase of predictors

Supplemental methods - Machine learning algorithms

Linear Regression

It fits a linear model with coefficients w = (w_1,…,w_p) to minimize the residual sum of squares between the actual responses in the dataset and the responses predicted by the linear approximation. p is the total number of demographics features.

Decision Tree [1]

It is a non-parametric supervised learning method, which aims to predict the value of a target variable by learning simple decision rules inferred from the data features. The model is obtained by recursively partitioning the data space and fitting a simple prediction model within each partition, where the partitioning can be represented graphically as a decision tree. Entropy is used to measure the quality of each split.

Support Vector Machine (SVM) [2]

It is a supervised learning machine for two-group classification problems. The main idea for SVM is to map the input feature vectors into a high-dimensional feature space through some non-linear mapping kernel function. Then, in this space a linear decision surface (or hyperplane) is constructed with special properties that ensure high generalization ability of the model to new data.

Random Forest [3]

A random forest is a meta-estimator consisting of a collection of tree-structured (decision tree) classifier on various sub-samples of the dataset. After a large number of trees is generated, each tree then casts a unit vote for determining the most popular class for a given new input sample.

AdaBoost [4]

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset. Then, it fits additional copies of the classifier on the same dataset, where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. The base classifier for AdaBoost is the Random Forest in this study.

Gradient Tree Boosting [5]

Gradient tree boosting or gradient boosted decision trees (GBDT) is a generation of boosting to arbitrary differentiable loss function. The model uses decision tree as the base classifier.

Feed-forward Neural Networks (FNN) [6]

It an artificial neural network model consisting of multi-layer perceptron (e.g., fully connected). It takes a feature vector as input and output a probability value to indicate which class to be assigned to for the input data. FNN optimizes the log-loss function using stochastic gradient descent.

Reference:

1. Loh, W.Y., 2011. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), pp.14-23.
2. Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), pp.1-27.
3. Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.
4. Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), pp.119-139.
5. Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.
6. Glorot, X. and Bengio, Y., 2010, March. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256). JMLR Workshop and Conference Proceedings.