Published: https://doi.org/10.1016/j.jcyt.2019.12.006

Predicting in vitro human mesenchymal stromal cell expansion based on individual donor characteristics using machine learning

Mohammad Mehrian[1,2], Toon Lambrechts[2,3], Marina Marechal[2,4], Frank P. Luyten[2,4], Ioannis Papantoniou[2,4,5] and Liesbet Geris[1,2,6]

[1] Biomechanics Research Unit, GIGA In Silico Medicine, University of Liege, CHU - BAT 34, Quartier Hopital, Avenue de l'Hopital 11, B-4000 Liege, Belgium.

[2] Prometheus, the Division of Skeletal Tissue Engineering, KU Leuven, Onderwijs en Navorsing 1 (+8), Herestraat 49 - PB813, B-3000, Leuven, Belgium.

[3] M3-BIORES, KU Leuven, Leuven, Onderwijs en Navorsing 1 (+8), Herestraat 49 -PB813, B-3000, Leuven, Belgium.

[4] Skeletal Biology and Engineering Research Center, KU Leuven, Leuven, Onderwijs en Navorsing 1 (+8), Herestraat 49 - PB813, B-3000, Leuven, Belgium.

[5] Institute of Chemical Engineering Science, Foundation of Research and Technology – Hellas (FORTH)

[6] Biomechanics Section, KU Leuven, Celestijnenlaan 300C (2419), B-3001, Leuven, Belgium.

Correspondence: Prof. Liesbet Geris

E-mail: Liesbet.geris@uliege.be

Abstract

Background: Human mesenchymal stromal cells (hMSCs) have become attractive candidates for advanced medical cell-based therapies. An *in vitro* expansion step is routinely used to reach the required clinical quantities. However, this is influenced by many variables including donor characteristics such as age and gender, and culture conditions, such as cell seeding density and available culture surface area. Computational modeling in general and machine learning in particular could play a significant role in deciphering the relationship between the individual donor characteristics and their growth dynamics. ***Methods:*** In this study, hMSCs obtained from 174 male and female donors, ages ranging from 3 to 64 with passage numbers ranging from 2 to 27 were studied. We applied a Random Forests (RF) technique to model the cell expansion procedure by predicting the population doubling time (PDT) for each passage, taking into account individual donor-related characteristics. ***Results:*** Using the RF model, the mean absolute error between model predictions and experimental results for the PDT in passage 1 to 4 is significantly lower compared to the errors obtained with theoretical estimates or historic data. Moreover, statistical analysis indicate that the PD and PDT in different age categories are significantly different, especially the youngest group (aged <10 yrs) compared to the other age groups. ***Discussion:*** In summary, we introduce a predictive computational model describing *in vitro* cell expansion dynamics based on individual donor characteristics, an approach that could greatly assist toward automation of a cell expansion culture process.

Keywords: Computational modeling; Donor characteristics; Human mesenchymal stromal cell; *In vitro* cell expansion; Random Forests.

Abbreviations: hMSC, Human mesenchymal stromal cells; **RF,** Random Forests; **PD,** population doubling; **PDT,** population doubling time; **hPDC,** Human periosteum derived stem cells; **P1,** passage 1; **ANN,** Artificial Neural Network; **SVM,** Support Vector Machines; **OOB,** Out-Of-Bag; **Ctime,** culture time; **CPD,** cumulative population doubling; **MAE,** mean absolute error; **MSE,** mean square error; **SE**, standard error

1 Introduction

Human mesenchymal stromal cells (hMSCs) are attractive candidates for a large number of cell-based therapies [1, 2]. In most cell-based therapies, a very large amount of cells is required which can vary from 10^6 to 10^8 cells depending on the application and treatment site [3]. However, a biopsy from a donor, be it a bone marrow of adipose tissue aspirate or a periosteal sample, only delivers a few thousand cells (per ml or mm^2, respectively). Therefore, *in vitro* cell expansion is an essential step in every cell-based therapy, which should be preferably fast but ultimately cost-effective [4].

Translating the cell expansion process towards clinical implementation requires a tight quality control of the process, including the preservation of cell identity and potency, with – where possible – removal of sources of variability. The variation in any biological product is mainly attributed to four factors: raw materials (including consumables), operational inputs (measurements, methods, personnel, equipment), environmental factors (e.g., change in room temperature within normal range), and biological variability inherent to living cells [5]. Inherent variability coming from the cells is a major source of variation in the autologous cell expansion process, where natural factors such as the age and gender of the donor have been shown to influence the expansion characteristics such as population doubling time (PDT) [6, 7]. Predicting the PDT for such a variable system can be challenging [3] and, in current practice, is mostly based on historically obtained average PDTs, complemented with visual inspection of the degree of confluency.

Various studies have been reported in the literature that investigate the link between donor characteristics and cellular behavior. Heathman et al. [8], used bone marrow derived stem cells from five male and female donors and measured the cumulative population doubling over five passages and discussed the implications of the measurable input variation such as the metabolite production rate and nutrient consumption rate on the manufacturing of cell-based therapies. Li et al. [9], studied the proliferation capacity of 17 samples of MSCs isolated from human bone marrow where an increase in population doubling time and decrease in the number of colonies formed was observed in older donors. Choi et al. [10] investigated how the morphology, cell-surface markers, proliferation potential and differentiation capacity of tonsil-derived MSCs from 20 donors aged from 5 to 54 years were affected by the donor age, long-term passage, and cryopreservation. Overall, these studies cover relatively small sets of donors, limiting the array of techniques used to analyze the obtained data to descriptive statistics mostly.

Machine learning is an area of artificial intelligence that uses statistical techniques to unveil insights and predictive relations in large data sets. Some examples of machine learning techniques include Artificial Neural Network (ANN), Support Vector Machines (SVM) and Random Forests (RF), each having their specific theoretical underpinnings and (dis)advantages. There is an increasing number of applications of machine learning in the biomedical context and only a few examples of these studies will be mentioned here. Furey et al. [11] used support vector machine classification for separating malignant from healthy tissue using microarray expression data. Monteiro et al. [12] used machine learning techniques to improve the prediction of functional outcome in ischemic stroke patients. Shaikhina et al. [13] used Random Forests models to predict the outcome in antibody incompatible kidney transplantation. Random Forests is a machine learning algorithm that is often applied in classification and regression problems [14]. It is frequently used due to its unique advantages

in dealing with small sample size, high-dimensional feature space, and complex data structures [15, 16].

In this study, we will apply machine learning to the data obtained from our periosteal hMSC donor bank containing 174 patient samples. We will start by explaining the experimental set-up used to collect the donor information followed by simple statistical analysis of the donor information. Subsequently, we will apply a Random Forests algorithm to obtain a predictive model for the cell expansion procedure, by predicting the PDT per passage. We will compare these predictions with the commonly used predictors for PDT, being the theoretical and historical means. To the best of our knowledge, this study is the first to apply machine learning on data derived from a sufficiently large MSC donor bank in order to predict the patient-specific *in vitro* cell expansion process.

2 Materials and Methods

2.1 Experimental set-up

2.1.1 Human periosteum derived stem cell culture

The cells used in this study are human periosteum derived stem cells (hPDC). These cells have been extensively studied by our lab and others, and have been shown to exhibit MSC features. Regardless of donor age and species of origin, these cells were shown to be clonogenic, could be expanded extensively in monolayer, displayed long telomeres and expressed typical MSC markers. Furthermore, under specific conditions, both parental and single-cell-derived clonal cell populations were able to differentiate into chondrocyte, osteoblast, adipocyte and skeletal myocyte lineages in vitro and in vivo [17, 18].

hPDCs were isolated from periosteal tissue biopsies from the tibia, obtained from patients undergoing orthopedic surgery to treat a non-union, perform limb lengthening or perform tibial varus/valgus correction. All procedures were approved by the ethics committee for Human

Medical Research (KU Leuven) and explicit patient (or parental) consent was obtained. In this study, donors having a known genetic defect impacting the skeleton were excluded.

The cells were expanded in monolayer in culture medium consisting of high glucose Dulbecco's modified Eagle's medium (Invitrogen, Merelbeke, Belgium) supplemented with 10% irradiated fetal bovine serum (iFBS; Gibco,Merelbeke, Belgium), 1% sodium pyruvate (Invitrogen) and 1% antibiotic–antimycotic (100 units/ml penicillin, 100 mg/ml streptomycin, and 0.25 mg/ml amphotericin B; Invitrogen) as described by Eyckmans and Luyten [18]). -

2.1.2 Cell culture metrics

The expansion of cells was performed in multiple passages of 2D culturing. In the first phase (passage 0 or P0), the cells from an enzymatically digested biopsy were relocated from a 6-well plate to a larger plastic surface and covered with growth medium. These plastic surfaces were tissue flasks with a surface of 25 cm^2, 75 cm^2 or 175 cm^2, depending on the number of cells available for seeding. The initial seeding density depended on the size of the sample obtained from the surgery and was therefore not the same for every donor. However, for all subsequent passages, a seeding density of 5700 cells/cm^2 was used throughout the expansion processes. Upon reaching near-confluence by visual inspection, cells were released from the substrate using trypsinization and a part of the cells (or all the cells) from the previous subculture were re-plated onto a new and larger plastic culture surface and covered with fresh growth medium, signifying the start of passage 1 (P1). Every subsequent passage started with a redistribution of cells over multiple new growth surfaces. This procedure was continued until the desired amount of cells was reached. The initial (and main) purpose of the hPDC bank was to be able to perform tissue engineering experiments with a clinically relevant cell source and to be able to investigate the effect of donor variation on said tissue engineering experiments. This explains the heterogeneity of the available information as cells were not routinely cultured up to senescence to record a full growth curve. Full growth curves were only recorded for a

limited number of donors, unrelated to the cells' behavior. In the data set used for this study, data from 174 male and female donors was included, ages ranging from 3 to 64 with passage numbers ranging from 2 to 27. For each donor only the age, gender and absence of known genetic disorders was available, in compliance with the GDPR standard.

2.1.3 Population doubling (PD) and population doubling time (PDT)

The number of population doublings is the number of times the population of cells doubles, which can be expressed per passage or as a cumulative number for the total expansion process. It is calculated as follows.

$$PD = log_2^{(\frac{Number\ of\ harvested\ cells}{Number\ of\ seeded\ cells})} \quad (1)$$

Another way to quantify the expansion of cells is by the population doubling time, which is the time needed for one population doubling. It is calculated as follows.

$$PDT = \frac{Culture\ time}{PD} \quad (2)$$

In every cell-based therapy, the population doubling time is of great importance since it is the main factor influencing the length of the expansion process and therefore a major cost driver. For example, for autologous therapies, the operators need to know how long it will take until they get the required amount of cells and, therefore, when the physicians can administer the therapy.

2.1.4 Statistical Analysis

Comparisons of PD or PDT in function of age or passage number were performed using one-way ANOVA tests in MATLAB (MathWorks®). A *p-value* < 0.05 was considered statistically significant.

2.2 Predicting the PDT

2.2.1 Rule of thumb estimates for PDT

According to the current practice in our lab, and most other labs, the PDT is estimated using a rule of thumb that is based either on theoretical arguments (gold standard) or historical observations, and does not take into account the individual donor characteristics. This gold standard states that adding one million HPDCs to a T175 flask, will lead to the harvest of about three million cells 7 days later. Therefore, the PD and PDT for every donor and in every passage would be 1.58 and 4.41 days respectively. Historic observations allow to make a calculation of the mean PD per passage over all donors in the data set, which is more interesting than having a single value for all passages. Indeed, it has repeatedly been shown that the cell growth rate slows down with increasing culture time [19]. This leads to mean values for PD and PDT in the following ranges: 2.55 (P1) – 0.52 (P20) and 3.23 (P1) – 28.74 (P20) days respectively.

2.2.2 Supervised learning methods and Random Forests (RF)

In supervised machine learning methods, the available data is separated in a training set and a test set. The training set is used to learn the hidden structure of the system by mapping the inputs to the outputs. The trained network therefore allows to predict a specific outcome based on a given input. This network is used subsequently to predict the output of the system for the unseen conditions in the test set [20, 21]. There are numerous studies that have applied supervised learning methods in a wide range of areas [22-24]. In this study we use the Random Forests (RF) technique for predicting the PDT.

Decision trees are tree-like graphs applied to both classification and regression problems. Decision trees can be displayed graphically which makes them very easy to interpret even by non-experts, but they suffer from high variance where a small change in the data can cause a large variation in the estimated output, making them non-robust [25]. ‘Bagging’ of

classification trees is an ensemble learning method which can overcome the problems of decision trees by constructing regression trees using bootstrapped training sets. The output is achieved by taking the average of the resulting predictions [26]. In 2001, Breiman proposed the RF algorithm which introduces an improvement over bagged trees by adding an additional layer of randomness to bagging that decorrelates the trees [14]. In bagging, each time a split in a tree occurs, the best split among all predictors is considered whereas in RF, each node is split using the best among a random subset of variables. Typically at each split, approximately a square root of the total number of predictors is considered [25]. Therefore, the RF algorithm only has two parameters to vary which are the number of allowed predictors at each split (*mtry*) and the total number of trees used in the forest (*ntree*). Usually the tree performance is not very sensitive to the value of these two parameters [27]. Additional information on the aforementioned concepts, as well as a graphical representation of the RF technique is provided in supplementary materials.

An important feature of RF technique is the Out-Of-Bag (OOB) error [28]. Each bagged tree uses around two-third of the observations. The remaining observations that are not used to fit a given bagged tree are specified as the OOB observations. Since the output for each observation is predicted using only the trees that were not fit using that observation, the OOB error is considered as a credible estimate of the test error for the bagged model. The OOB error of the RF is obtained by taking the average error of the observations from the dataset using the OOB trees. Therefore, this built-in internal validation causes the error estimation to be less optimistic and is generally considered as a reliable estimator of the expected error for independent data [25]. Furthermore, this feature helps RF to avoid overfitting [14, 16].

In this study, a RF model for predicting the PDT in P1 to P4 was created in the software R, using its *randomForest* package. For predicting the PDT of P1 (referred to as PDT1 in the remainder of the paper), the input parameters were the (i) age and (ii) gender of the donor, (iii)

the available culture surface area used in the passage for P1, (iv) the number of cells seeded at the beginning of the passage in P1 and (v) the culture time (*Ctime*) of the previous (P0) passage. For predicting the PDT2, PDT3 and PDT4, the PDT and the *Ctime* of all previous passages were also added to the input parameters to further improve the performance of the model. The prediction results of the RF method were compared with the rule of thumb methods' PDT estimates.

3 Results

3.1 Analysis of the donor population and cell growth characteristics

In the period 2012-2018, 174 donors were added to the cell bank. Given the variability in the amount of cells extracted from the biopsy, information on population doubling times at passage 0 (P0) is not available for all donors. Therefore, in this study, we did not include a model for P0 and all model predictions start from P1. An overview of all 174 donors based on the number of donors in each passage (from P1 to P20) is shown in Figure 1A. Figure 1B shows the number of donors in different age categories in the entire donor set and for the selection used in the model (cfr infra). Out of the 174 donors, only 4 donors were cultured for more than 17 passages. Figure 1C shows the cumulative population doubling (CPD) of these 4 donors over time, labeled by their respective ages at harvest.

In order to investigate if the age of donors has any effect on the PD and PDT in different passages, we have categorized donors into six different age groups. Groups one to five comprise donors with the ages of 0 to 50 years, divided in 10 year intervals. Group six are the oldest donors with ages over 50 years. Figures 2 and 3 show the PD and the PDT from P1 to P4 for the aforementioned age categories respectively.

As is shown, the age of the donor has an important effect on the amount of PD and PDT where younger donors, on average, have higher PD and lower PDT. Significant differences were

observed between the PD and PDT of the first group aged less than 10 years with the rest of the groups, except for PD1 where the significant difference is only between the first and the second age group. No significant differences were observed between the donors in the middle age categories (20 to 50 years).

To better evaluate the PD and PDT over the entire culture period, the mean and standard deviation of PD and PDT in each passage is calculated. Note that, the number of donors in each passage used for calculating the mean and standard error are different, as shown in Figure 1A.

As is shown in Figure 4A, the first few passages have the highest PD and the average PD is decreasing as the number of passages are increasing. The average PDT seems to slightly increase over different passages (Figure 4B) but no significant differences could be detected due to the large variation in the data set. Figure 4C shows the mean and standard deviation of the CPD of all donors from P1 to P6 divided in their respective age categories. Data is shown up until P6 as there are only a few donors (~10% of the total donor set) that go beyond P6. The average amount of CPD in all passages is higher for younger donors and a clear correlation between CPD and age of the donor is observed. The youngest group (aged 10 years or less), have the highest CPD in all passages.

3.2 Predictive modeling of population doubling time

Out of 174 donors, 141 donors were cultured at least up to P4. The 33 donors for which the growth curve up to P4 was not available were excluded from the data set. Additionally, for a limited amount of donors in certain passages, the amount of harvested cells was only slightly higher – or even lower – than the amount of seeded cells, leading to a very large– or negative – value for the PDT. For example, for a specific donor in P3, after 7 days, we have harvested 940.000 cells while we started with 10^6 seeded cells. This would lead to a value for PD and PDT of *-0.0893* and *-78.41 days* respectively. For other donors, certain passage's culture time

was prolonged beyond the usual 1-week time point due to slow cell growth. Out of the 141 remaining donors, 10 donors had at least 1 PDT greater than 10 days or 1 negative PDT. As the PDT values for previous and subsequent passages for those 10 donors were in the normal ranges, these particular passages were considered to be outliers and the donors were omitted from the data set. This omission was performed manually in this study, but could easily be included in the model as a pre-model quality check using the same criteria as for the manual curation step. After this step, 131 donors remained to build a predictive model for PDT in P1 to P4. This exclusion criteria can be used during the actual expansion process to delineate the context of use of the RF model (i.e. in which conditions the model applies). The number of donors in each age category for the 131 donors used in the model is shown in figure 1B. Donors from all age categories were present in the data set, therefore, the model trained on this data is be applicable to all age groups.

An RF model was constructed for each passage separately using 70% of the donors (91 donors) for training the RF model and 30% of the donors (40 donors) for testing the performance of the model after its construction. To optimize the performance of the RF model, the two main variables of the model, *mtry* and *ntree,* were tuned. An *mtry* value of 1 resulted in the lowest OOB error. As mentioned earlier, the RF technique is not very sensitive to the number of trees and we did not observe a noticeable difference in the performance of the model using different tree numbers. Therefore, the RF model using 1000 trees with one variable randomly sampled at each split was identified to perform best. The OOB error for different *mtry* values and the MAE of the prediction for different *ntree* values are provided in the supplementary material.

The results of the RF model for predicting the PDT in P1 to P4 are shown in Figure 5 with model predictions for the donors in the test set that was unseen by the trained RF model (red points), and the experimental values (blue points) in each passage. As the RF model gains more knowledge of the underlying system by including more information from previous passages

such as their PDT and *Ctime*, the predictions in higher passages (P2, P3 and P4) improve compared to previous passage. Quantification of the mismatch between model predictions and experimental observations shows that the mean absolute error (MAE) is consistently decreasing from 0.87 days for PDT1 to 0.7 days for PDT4. The variable importance plots (panels on the right) show the percentage increase in the mean square error (MSE), in case each of the input variables are permuted. For example in PDT1, the culture time of the previous passage (P0) plays the most important role in predicting the PDT1 followed by the age of the donor. For each passage, the importance of each variable having the most effect on the PDT is shown. In all passages, the PDT of the previous passage is the most influential factor for predicting the PDT of the current passage.

Figure 6 shows the comparison between RF model predictions and the predictions obtained using the rule of thumb estimates, both for the gold standard and the mean historical data. The comparison is made for all 141 donors (cultured at least up to P4), and for the case where the 10 problematic donors having PDT<0 or PDT>10 have been omitted.

For the gold standard method, the average PDT is calculated based on theoretical considerations, resulting in the PDT of 4.41 days for every passage. For the mean historical data method, the whole data set is used to calculate the average PDT in each passage. For all methods in all passages, excluding the problematic donors based on the aforementioned exclusion criterion, the MAE and the standard error decrease and the performance improves. For the mean historical data method, with or without the donor omission, the mean and standard error are smaller than for the gold standard method (except for P3) as we calculate the average PDT over each passage resulting in a more accurate estimate of PDT. The best performance is achieved using the RF model after donor exclusion (light pink).

To further assess the accuracy of the RF model and particularly its dependency on the choice of training vs test set, the RF model was trained 100 times with random splitting of data in training (70%) and test (30%) data sets. Each time the MAE of the predicted PDT for P1 to P4 is calculated and ultimately the average MAE of the 100 RF model repeats can be calculated for each passage. For P1 to P4 the mean MAE is 0.99, 0.87, 0.80 and 1 day respectively with the mean standard deviation of 0.1, 0.08, 0.09 and 0.11 (Figure 6, yellow). As expected, using different random splits for the training and test data set did not affect the outcome as in the RF technique there is a built-in validation process, as explained in previous sections.

4 Discussion

In this study, we have presented the application of a supervised learning method (Random Forests) for predicting the PDT for a 2D flask-based in vitro cell expansion procedure. Based on the statistical analysis, we showed that the PD and PDT in different age categories are significantly different, especially the youngest group (aged <10) compared to the other age groups. The accuracy of the RF model in predicting the PDT was better (lower prediction error) than the rule of thumb methods in all of the four passages. According to RF model results, the age of the donor and the PDT of the previous passages are the most influential factors for predicting the PDT values of different passages.

There are a number of studies on *in vitro* cell expansion in which the effect of different factors such as age of the donor on cell growth has been investigated [29-32]. But in all of these studies, the number of donors being investigated was limited (e.g., 53 individuals in [33]) and the focus was more on the biological characteristics of the cell expansion procedure such as the osteogenic differentiation abilities and the gene expression of the donors related to their age and culture conditions [33-35]. To the best of our knowledge there is no previous study using modeling to predict the cell growth in different passages of the cell expansion process. In this

study we have predicted the PDT per passage rather than the cumulative PDT at the end of the culture period, as the former allows tracking the growth rate as well as the time in each passage and has a direct impact on the cell culture process.

One of the main limitations in the computational modeling of the cell expansion procedure is the lack of donor data. The highest number of donors presented in cell expansion literature consisted of 53 individuals, 25 female, 28 male aged between 13 to 80 years [33], an amount that is not sufficient to build a reliable model using machine learning techniques. Even for the larger set used in this study we had to use the RF technique which is the most efficient method, able to deal with small sample sizes [15, 16]. RF uses a bootstrapping technique to fully exploit the training data set. Note that we also explored the use of neural networks to model the cell expansion process (results not shown), but with the size of the data set used in this study, the MAE obtained in each passage was very high compared to RF results, making the method unfit for this application. In the developed model, we accounted for the risk of overfitting which refers to the fact that a model that adapts too closely to the learning samples will not be able to discover all the patterns in the training data set and thus cannot generalize well for the test set. The theoretical result shown in Breiman [26] supports the fact that ensemble methods such as Random Forests do not overfit with an increase in the number of trees. However, it has been shown that that they might overfit due to other reasons such as the depth of the grown trees [36]. In this study, we have tuned the *mtry* value which introduces more randomness to the grown trees by making them more diverse and as a result reduces the risk of overfitting. Moreover, the average MAE obtained by training the model 100 times (yellow bar in Fig. 5) shows very similar error values with the built RF model in all of the four passages (purple bar in Fig. 5) which is a an indication of the robustness and reliability of the built model.

Other machine learning techniques such as support vector machines [37] and boosting [38] could be applied to our problem. The performance of these two methods is very close to the

Random Forests we used in this study. For example in Wu et al. [39], authors compared the performance of several types of statistical methods such as linear discriminant analysis, k - nearest neighbor classifier, bagging and boosting classification trees, support vector machines and random forest for the classification of cancer. They have shown that random forests technique outperforms other competitors, although in Konig et al. [40], SVM slightly outperforms Random Forests for patient-centered yes/no prognosis. All in all, considering the ease of application, most of the machine learning techniques need extra tuning to deliver their best performance, whereas Random Forests is believed to be among the best performing methods even without extra tuning [41].

An important aspect of modeling is the biological interpretation of the results. The developed RF model is capable of identifying the most important input variables affecting the predicted PDT values in each passage. For example, it is shown that age of the donor plays the most important role in cell growth for the first two passages, which is in corroboration with other studies in the literature [32, 33]. Marędziak et al. [42] compared the PD and PDT in 28 patients aged from 22 to 77 years where younger donors showed significantly higher PD and lower PDT values. In Fig. 3A we showed that the PD decreases with the increase in passage numbers. Other studies have reported the same trend for PD over different passages. In Zaim et al. [7] authors showed that long term passage affects both the morphology and proliferation of hMSCs in all ages where the increase in passage number led to a decrease in proliferation rate of the cells. According to the model results, the previous PDT becomes more important than the age of the donor as we go toward higher passages.

Additionally, almost in all passages, the gender of the donor is the least important feature of the input space and the cell growth is deemed to be independent of donor gender. As in most

published studies male and female patients are grouped together, there is little literature specifically studying the effect of gender on hMSC proliferation. Furthermore, mixed results were published on the effect of gender on hMSC proliferation, as reviewed by Fossett and Khan [43].

Although the seeding density in our experiments was almost constant (5700 cells/cm^2) in all passages after P0, it was taken along in the creation of the RF model. Consistent literature reports have shown that lower seeding densities will result in faster proliferation rates both in T-flasks and three dimensional scaffolds [44, 45].

This model was developed for manual cell culture where information of the process was limited to measurements of cell number between passages. The implementation of such approaches in bioreactors is possible and could in fact result in more accurate predictions since the robustness of the process of the bioreactor as compared to manual expansion might enhance this correlation. For hPDCs we have already shown that this applies since process performance substantially increases in bioreactors [46]. In addition, data regarding cell growth could be either monitored online through imaging [47] or other sensors such as oxygen and lactate data [48], that once validated can provide more dense information throughout cell expansion [49] and not only during passaging. Especially with the development of fully controlled and sensor-embedded bioreactor systems [50] such approaches could be embedded in control strategies, 'correcting' for donor dependent variation.

Given large enough data-sets our approach could be expanded to iPS cells or allogeneic cells and could be used for allogeneic and hence scale-up scenarios. However, in this work we focus on the autologous application where donor-to-donor variability is an essential source of variability in the starting material, as shown by our analysis (Figure 1-3) but also by additional

work in bioreactors [51]. Donor-related characteristics is mostly indicative of scale-out scenarios and knowledge extracted through our methods can inform the design of expansion processed or prevention of batch failures early on.

In Figure 6, the performance of the Gold standard and mean historical method are close to the RF model in the first two passages and the MAE is less than 2 days which explains why this rule of thumb has been used successfully to date. When sufficient in-house data is available, the mean historic average provides a good estimate, even though in practice the theoretical estimate often remains the standard. However, with increasing passage number the MAE of the RF model drops significantly compared to the standard estimates. When moving towards more clinical applications which require large amounts of cells and hence extensive passaging, the advantages of the RF model will become increasingly relevant.

One of the main limitations of the built RF model is the inability to predict the PDT of the 10 excluded donors having negative or very little growth during a particular passage of the expansion process, resulting in negative or very high PDT values. As previously mentioned, the main source for these PDT values might be the operational inputs rather than the inherent biological characteristics as in these problematic donors the PD or PDT values of the preceding and subsequent passages were normal. In the current study, taking along these 10 donors resulted in a poorer model. Yet, the exclusion criteria used to construct the model can also be used during culture to delineate its use. The exclusion criteria can easily be implemented in the model itself as an initial quality check. For donors that fail this quality check, the model cannot be applied to reliably to predict the next passage outcome and the rule of thumb estimate along with repeated visual inspection is required. Once the expansion behavior meets again the quality criteria, further passages can be predicted again.

For the first passage, the initial seeding density (P0) could not be used as that information was not available for all donors. This means that the model is not able to incorporate the real in vitro age of the cells at the first passage, which is something that should be remedied in the future.

Another limitation of the model is that using the current dataset we are unable to go beyond passage P4 for predicting PDT due to the lack of sufficient donor data at higher passages. However, in the future, another version of the RF model could be built where the predicted PDT of the previous passage is added as an input for the next passage, rather than the actual experimental values of the PDT that we used in the model. In the case of having satisfactory results (low prediction error in passage P2, P3 and P4), we can predict the PDT up to passage P4 in advance. It should be noted that there is no limitation in terms of the type of the MSCs used in the model. Given the appropriate donor and culture information for sufficient donors, the model is capable of predicting that particular cell expansion process. Finally, using the same methodology, we could have built a model that predicts the amount of PD of the cells. Predicting the PDT is believed to be more informative than predicting the PD, as in PDT, both the PD and the culture time are included.

In conclusion, the model presented in this study could be used as a tool assisting cell expansion operators in predicting the time it takes to reach a certain amount of cells based on the characteristics of the donor using well-defined culture conditions. In our model, the overall prediction error is significantly different from the rule of thumb methods in all passages (especially for the model with 131 donors). Accordingly, the presented model shows great potential in employing data driven modeling for the characterization and prediction of *in vitro* cell expansion .

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 279100, the European Union's Horizon 2020 Programme (FP/2014-2020)/ERC CoG 772418 and the Belgian National Fund for Scientific Research (FNRS) grant FRFC 2.4564.12 and T025413F. I.P. was funded by Fonds Wetenschappelijk Onderzoek (FWO) Fellowship (project n. 12O7916N). T.L was funded by VLAIO (project n. HBC.2016.0629). We thank Kathleen Bosmans, Carla Geeroms, Loes Vanhoudt, Stijn Bellinckx, Lennert Ceysens and Melanie Van den Broeck for their contribution to laboratory experiments. This work is part of Prometheus, the Leuven Research and Development Division of Skeletal Tissue Engineering of the KU Leuven: www.kuleuven.be/prometheus.

Disclosure of Interest

The authors declare no financial or commercial conflict of interest.

5 References

[1] Caplan, A. I., Adult mesenchymal stem cells for tissue engineering versus regenerative medicine. *Journal of cellular physiology* 2007, *213*, 341-347.
[2] Pittenger, M. F., Mackay, A. M., Beck, S. C., Jaiswal, R. K.*, et al.*, Multilineage potential of adult human mesenchymal stem cells. *science* 1999, *284*, 143-147.
[3] Heathman, T. R., Nienow, A. W., McCall, M. J., Coopman, K.*, et al.*, The translation of cell-based therapies: clinical landscape and manufacturing challenges. *Regenerative medicine* 2015, *10*, 49-64.
[4] Weiss, L. E., Junkers, S. N., Chen, M., Yin, Z.*, et al.*, An engineered approach to stem cell culture: Automating the decision process for real-time adaptive subculture of stem cells. *PLoS One* 2011, *6*, e27672.
[5] Shimoni, Y., Forsyth, T., Srinivasan, V., Szeto, R., Reducing Variability in Cell-Specific Productivity in Perfusion Culture: A Case Study. *The american chemical society* 2017, *253*.
[6] Fossett, E., Khan, W. S., Longo, U. G., Smitham, P. J., Effect of age and gender on cell proliferation and cell surface characterization of synovial fat pad derived mesenchymal stem cells. *Journal of Orthopaedic Research* 2012, *30*, 1013-1018.
[7] Zaim, M., Karaman, S., Cetin, G., Isik, S., Donor age and long-term culture affect differentiation and proliferation of human bone marrow mesenchymal stem cells. *Annals of Hematology* 2012, *91*, 1175-1186.

[8] Heathman, T. R., Rafiq, Q. A., Chan, A. K., Coopman, K.*, et al.*, Characterization of human mesenchymal stem cells from multiple donors and the implications for large scale bioprocess development. *Biochemical engineering journal* 2016, *108*, 14-23.
[9] Li, Y., Charif, N., Mainard, D., Bensoussan, D.*, et al.*, Donor's age dependent proliferation decrease of human bone marrow mesenchymal stem cells is linked to diminished clonogenicity. *Bio-medical materials and engineering* 2014, *24*, 47-52.
[10] Choi, J.-S., Lee, B.-J., Park, H.-Y., Song, J.-S.*, et al.*, Effects of donor age, long-term passage culture, and cryopreservation on tonsil-derived mesenchymal stem cells. *Cellular Physiology and Biochemistry* 2015, *36*, 85-99.
[11] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W.*, et al.*, Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000, *16*, 906-914.
[12] Monteiro, M. A. B., Fonseca, A. C., Freitas, A. T., e Melo, T. P.*, et al.*, Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2018.
[13] Shaikhina, T., Lowe, D., Daga, S., Briggs, D.*, et al.*, Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control* 2017.
[14] Breiman, L., Random forests. *Machine learning* 2001, *45*, 5-32.
[15] Qi, Y., Random forest for bioinformatics, *Ensemble machine learning*, Springer 2012, pp. 307-323.
[16] Yang, P., Hwa Yang, Y., B Zhou, B., Y Zomaya, A., A review of ensemble methods in bioinformatics. *Current Bioinformatics* 2010, *5*, 296-308.
[17] De Bari, C., Dell'Accio, F., Vanlauwe, J., Eyckmans, J.*, et al.*, Mesenchymal multipotency of adult human periosteal cells demonstrated by single-cell lineage analysis. *Arthritis & Rheumatism* 2006, *54*, 1209-1221.
[18] Eyckmans, J., Luyten, F. P., Species specificity of ectopic bone formation using periosteum-derived mesenchymal progenitor cells. *Tissue engineering* 2006, *12*, 2203-2213.
[19] Chen, J., Du, X., Chen, Q., Xiang, C., Effects of donors' age and passage number on the biological characteristics of menstrual blood-derived stem cells. *International journal of clinical and experimental pathology* 2015, *8*, 14584.
[20] Mohri, M., Rostamizadeh, A., Talwalkar, A., *Foundations of machine learning*, MIT press 2012.
[21] Baldi, P., Deep Learning in Biomedical Data Science. *Annual Review of Biomedical Data Science* 2018, *1*, null.
[22] Shen, D., Wu, G., Suk, H.-I., Deep learning in medical image analysis. *Annual review of biomedical engineering* 2017, *19*, 221-248.
[23] LeCun, Y., Bengio, Y., Hinton, G., Deep learning. *nature* 2015, *521*, 436.
[24] Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., *Deep learning*, MIT press Cambridge 2016.
[25] James, G., Witten, D., Hastie, T., Tibshirani, R., *An introduction to statistical learning*, Springer.
[26] Breiman, L., Bagging predictors. *Machine learning* 1996, *24*, 123-140.
[27] Liaw, A., Wiener, M., Classification and regression by randomForest. *R news* 2002, *2*, 18-22.
[28] Mitchell, M. W., Bias of the Random Forest out-of-bag (OOB) error for certain input parameters. 2011.
[29] Detela, G., Bain, O. W., Kim, H. W., Williams, D. J.*, et al.*, Donor variability in growth kinetics of healthy hMSCs using manual processing: considerations for manufacture of cell therapies. *Biotechnology journal* 2018.

[30] Lambrechts, T., Sonnaert, M., Schrooten, J., Luyten, F. P.*, et al.*, Large-scale mesenchymal stem/stromal cell expansion: a visualization tool for bioprocess comparison. *Tissue Engineering Part B: Reviews* 2016, *22*, 485-498.
[31] Zuk, P. A., Zhu, M., Mizuno, H., Huang, J.*, et al.*, Multilineage cells from human adipose tissue: implications for cell-based therapies. *Tissue engineering* 2001, *7*, 211-228.
[32] Bonab, M. M., Alimoghaddam, K., Talebian, F., Ghaffari, S. H.*, et al.*, Aging of mesenchymal stem cell in vitro. *BMC cell biology* 2006, *7*, 14.
[33] Siegel, G., Kluba, T., Hermanutz-Klein, U., Bieback, K.*, et al.*, Phenotype, donor age and gender affect function of human bone marrow-derived mesenchymal stromal cells. *BMC medicine* 2013, *11*, 146.
[34] Basciano, L., Nemos, C., Foliguet, B., de Isla, N.*, et al.*, Long term culture of mesenchymal stem cells in hypoxia promotes a genetic program maintaining their undifferentiated and multipotent status. *BMC cell biology* 2011, *12*, 12.
[35] Kim, J., Kang, J. W., Park, J. H., Choi, Y.*, et al.*, Biological characterization of long-term cultured human mesenchymal stem cells. *Archives of pharmacal research* 2009, *32*, 117-126.
[36] Luellen, J. K., Shadish, W. R., Clark, M., Propensity scores: An introduction and experimental test. *Evaluation Review* 2005, *29*, 530-558.
[37] Vapnik, V., *The nature of statistical learning theory*, Springer science & business media 2013.
[38] Freund, Y., Schapire, R. E., A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 1997, *55*, 119-139.
[39] Wu, B., Abbott, T., Fishman, D., McMurray, W.*, et al.*, Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 2003, *19*, 1636-1643.
[40] Konig, I., Malley, J. D., Pajevic, S., Weimar, C.*, et al.*, Patient-centered yes/no prognosis using learning machines. *International journal of data mining and bioinformatics* 2008, *2*, 289-341.
[41] Caruana, R., Niculescu-Mizil, A., *Proceedings of the 23rd international conference on Machine learning*, ACM 2006, pp. 161-168.
[42] Marędziak, M., Marycz, K., Tomaszewski, K. A., Kornicka, K., Henry, B. M., The influence of aging on the regenerative potential of human adipose derived mesenchymal stem cells. *Stem cells international* 2016, *2016*.
[43] Fossett, E., Khan, W., Optimising human mesenchymal stem cell numbers for clinical application: a literature review. *Stem cells international* 2012, *2012*.
[44] Both, S. K., Muijsenberg, A. J. v. d., Blitterswijk, C. A. v., Boer, J. d., Bruijn, J. D. d., A rapid and efficient method for expansion of human mesenchymal stem cells. *Tissue engineering* 2007, *13*, 3-9.
[45] Lode, A., Bernhardt, A., Gelinsky, M., Cultivation of human bone marrow stromal cells on three-dimensional scaffolds of mineralized collagen: influence of seeding density on colonization, proliferation and osteogenic differentiation. *Journal of tissue engineering and regenerative medicine* 2008, *2*, 400-407.

[46] Lambrechts T, Papantoniou I, Viazzi S, Bovy T, Schrooten J, Luyten FP, Aerts JM. Evaluation of a monitored multiplate bioreactor for large-scale expansion of human periosteum derived stem cells for bone tissue engineering applications. Biochemical engineering journal. 2016 Apr 15;108:58-68.

[47] Viazzi S, Lambrechts T, Schrooten J, Papantoniou I, Aerts JM. Real-time characterisation of the harvesting process for adherent mesenchymal stem cell cultures based on on-line imaging and model-based monitoring. Biosystems Engineering. 2015 Oct 1;138:104-13.

[48] Lambrechts T, Papantoniou I, Sonnaert M, Schrooten J, Aerts JM. Model-based cell number quantification using online single-oxygen sensor data for tissue engineering perfusion bioreactors. Biotechnology and bioengineering. 2014 Oct;111(10):1982-92.

[49] de Bournonville S, Lambrechts T, Vanhulst J, Luyten FP, Papantoniou I, Geris L. Towards self-regulated bioprocessing: a compact benchtop bioreactor system for monitored and controlled 3D cell and tissue culture. Biotechnology journal. 2019 Apr 9:1800545.

[50] Das R, Roosloot R, van Pel M, Schepers K, Driessen M, Fibbe WE, de Bruijn JD, Roelofs H. Preparing for cell culture scale-out: establishing parity of bioreactor-and flask-expanded mesenchymal stromal cell cultures. Journal of translational medicine. 2019 Dec 1;17(1):241.

[51] Heathman TR, Rafiq QA, Chan AK, Coopman K, Nienow AW, Kara B, Hewitt CJ. Characterization of human mesenchymal stem cells from multiple donors and the implications for large scale bioprocess development. Biochemical engineering journal. 2016 Apr 15;108:14-23.

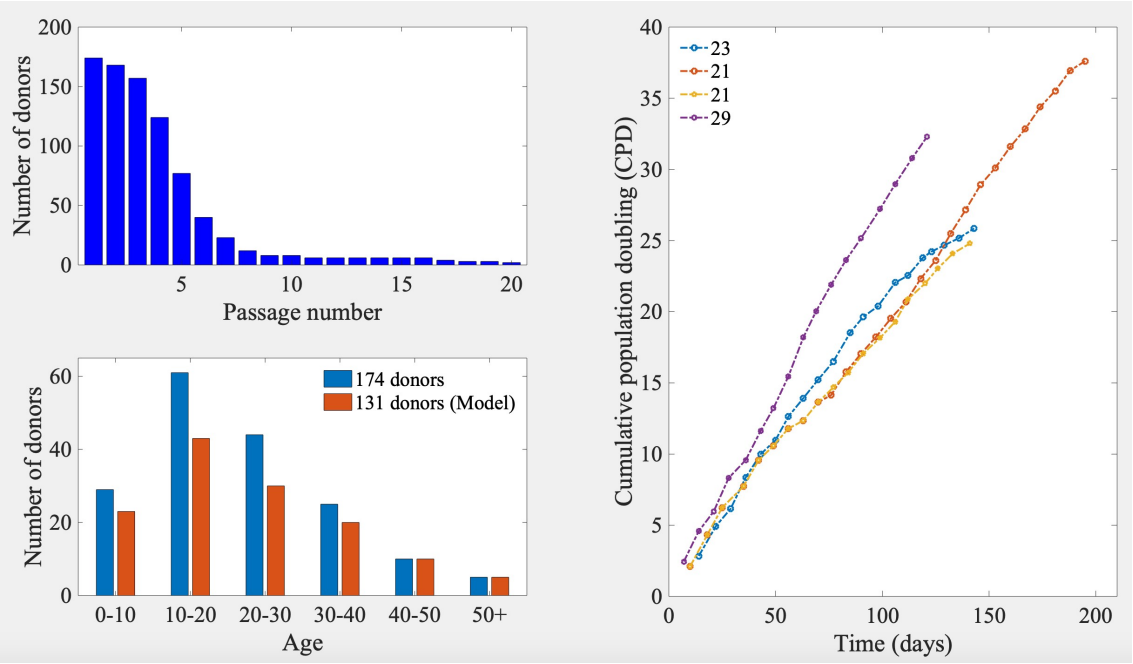

Figure 1. Overview of donor population. (A) Summary of all donors based on the number of donors in each passage (from P1 to P20). (B) The number of donors in different age categories for the full (blue) and the curated (red) data set. (C) The cumulative population doubling (CPD) of periosteal cells cultured for more than 20 passages, labeled by the donor's age at the moment of periost donation. (D) The number of donors used for the modeling section in different age categories.

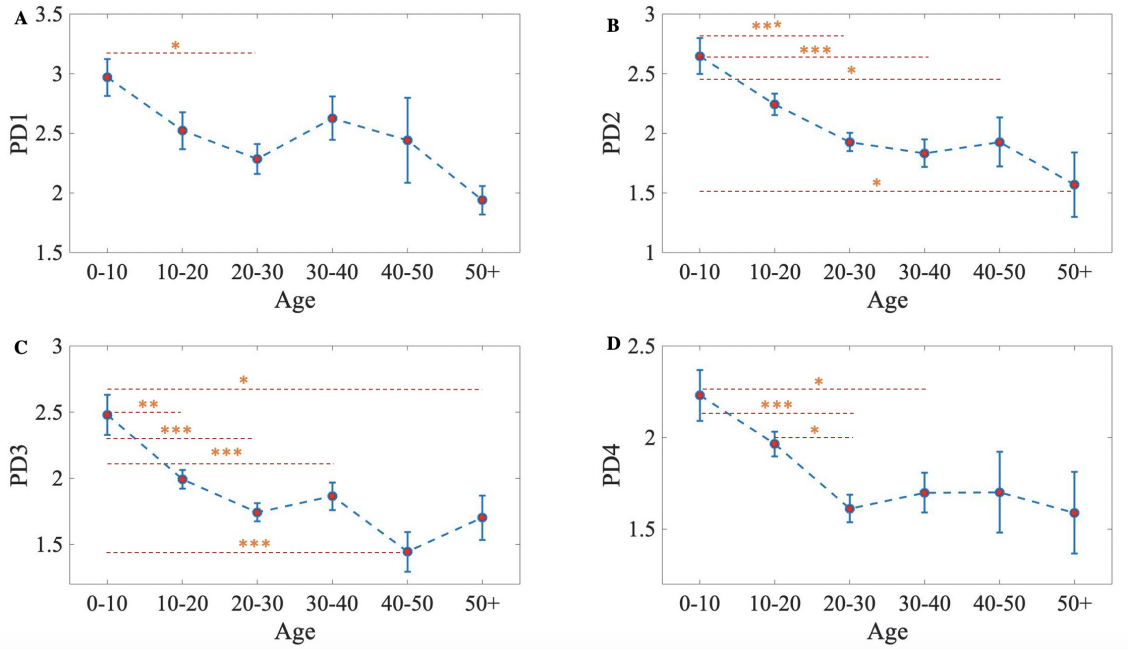

Figure 2. Representation of cell growth characteristics in function of age groups for P1-P4. The mean and standard error of population doubling (PD) for P1 to P4 in function of the age categories. *$p < 0.05$; **$p < 0.005$; ***$p < 0.0001$

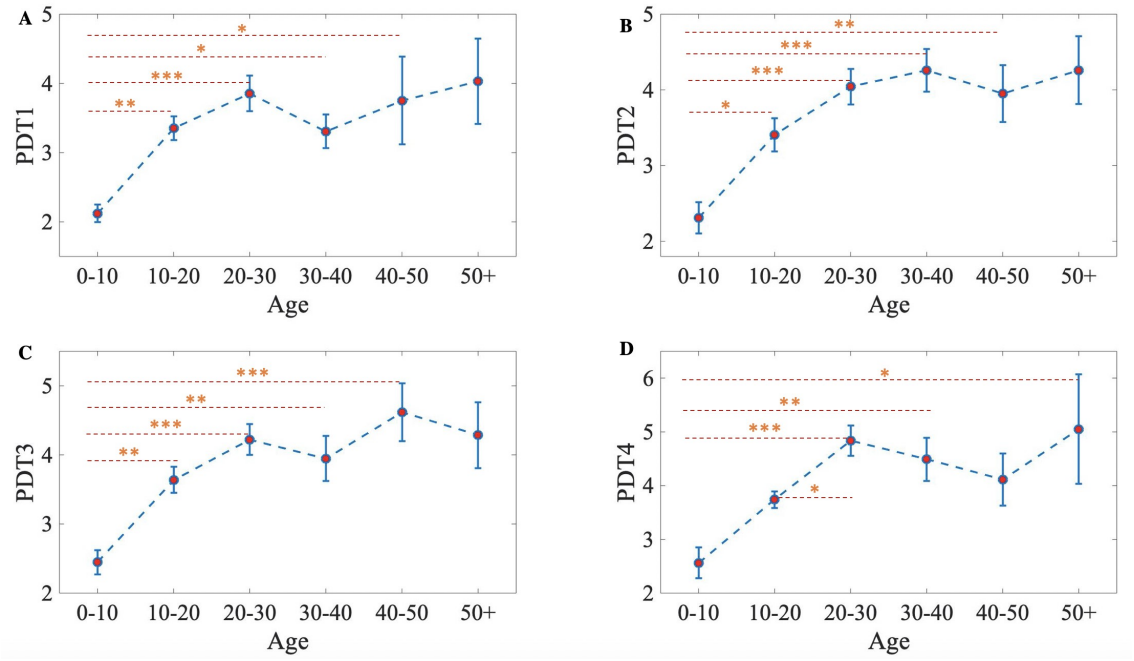

Figure 3. Representation of cell growth characteristics in function of age groups for P1-P4. The mean and standard error of population doubling time (PDT) for P1 to P4 in function of the age categories. *p < 0.05; **p < 0.005; ***p < 0.0001

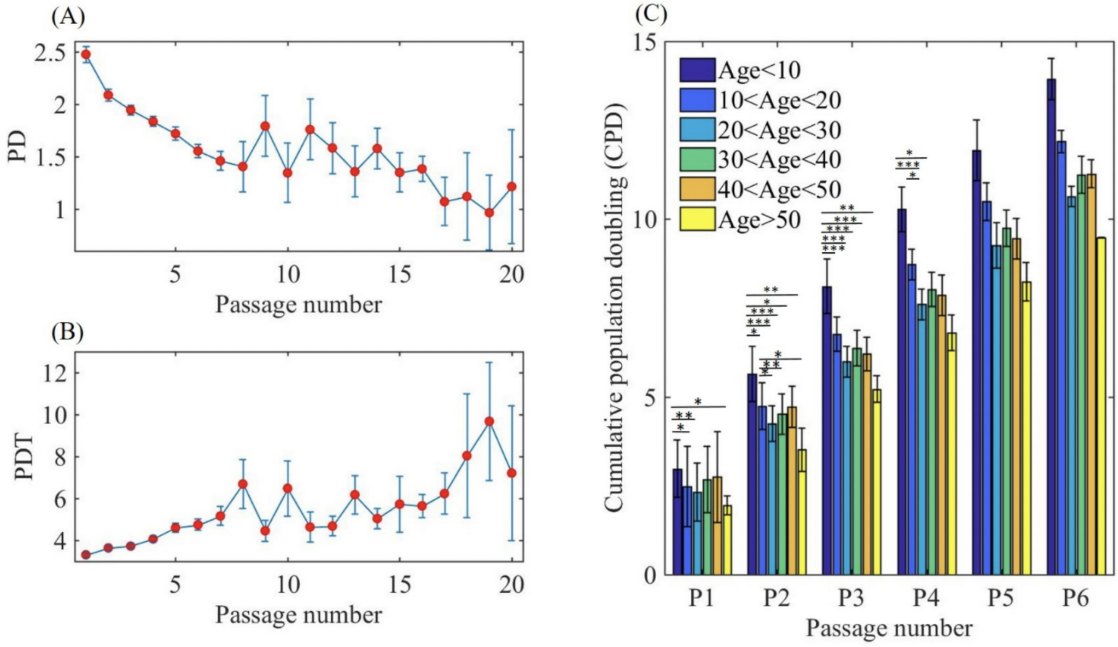

Figure 4. Analysis of cell growth characteristics in function of passage number and age. (A) The mean and standard error of population doubling (PD) in each passage. (B) The mean and standard error of the population doubling time (PDT) in each passage. (C) The mean and standard error of cumulative population doublings (CPD) of donors from P1 to P6 in different age categories. *$p < 0.05$; **$p < 0.005$; ***$p < 0.0001$

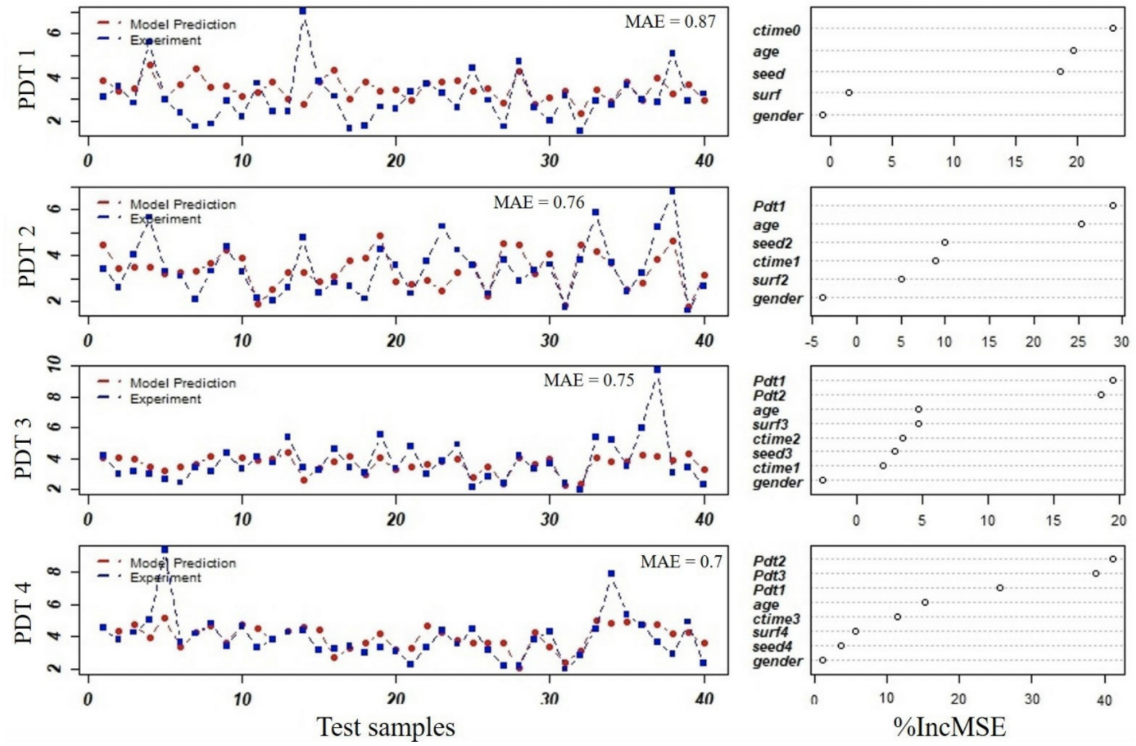

Figure 5. Comparison of model predictions and experimental observations, and the relative parameter importance. The results of the RF model for predicting the PDT in P1 to P4 (red dots) and the experimental observations (blue dots) for all 40 donors in the test set. The prediction's mean absolute error (MAE) is indicated in the upper right corner. The variable importance plot for each passage is shown on the right hand side, indicating the relative contribution of each parameter in the total increase in the mean square error (MSE) of prediction [%].

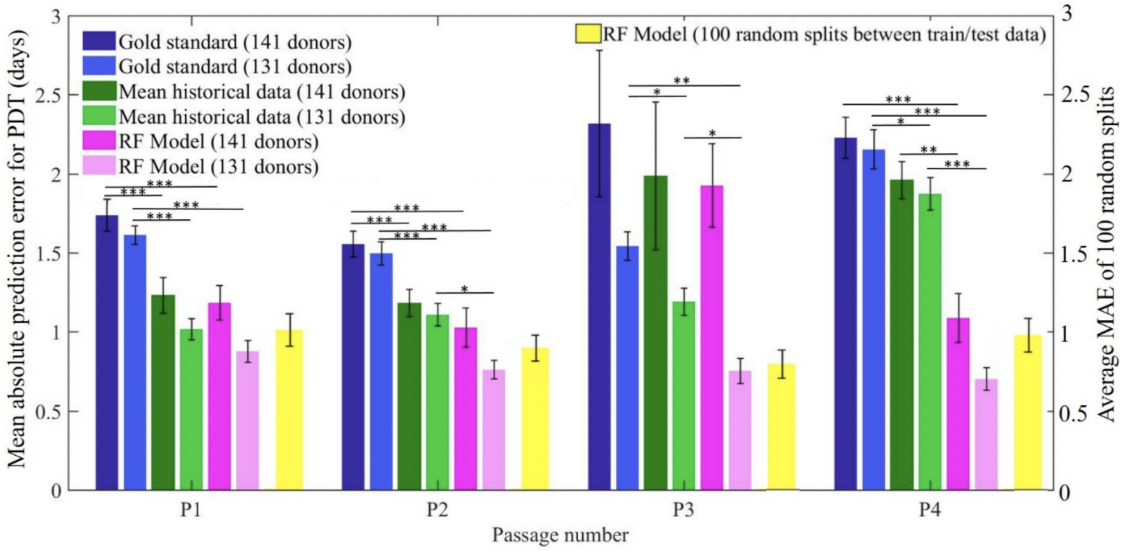

Figure 6. Mean Absolute Prediction error for rule of thumb models (gold standard and mean historic) and RF models (left axis). Errors are calculated for the full data set (141 donors) or the data set after omitting problematic donors (131 donors) and are shown as mean ± SE The right bar for each passage shows the average of the mean absolute error (MAE) of running the RF model 100 times using random splits between train and test data. *$p < 0.05$; **$p < 0.005$; ***$p < 0.0001$.