RESEARCH ARTICLE

Personalized pathology test for Cardio-vascular disease: Approximate Bayesian computation with discriminative summary statistics learning

Ritabrata Dutta [1]*, **Karim Zouaoui Boudjeltia** [2], **Christos Kotsalos** [3], **Alexandre Rousseau** [2], **Daniel Ribeiro de Sousa** [2], **Jean-Marc Desmet** [4], **Alain Van Meerhaeghe** [5], **Antonietta Mira** [6,7], **Bastien Chopard** [3]

1 University of Warwick, United Kingdom, **2** Laboratory of Experimental Medicine (ULB 222), Medicine Faculty, Université Libre de Bruxelles, ISPPC CHU de Charleroi, Charleroi, Belgium, **3** University of Geneva, Geneva, Switzerland, **4** Nephrology Department, ISPPC CHU de Charleroi, Charleroi, Belgium, **5** Pneumology Department, ISPPC CHU de Charleroi, Charleroi, Belgium, **6** Università della Svizzera italiana, Lugano, Switzerland, **7** University of Insubria, Varese, Italy

* Ritabrata.Dutta@warwick.ac.uk

Abstract

Cardio/cerebrovascular diseases (CVD) have become one of the major health issue in our societies. But recent studies show that the present pathology tests to detect CVD are ineffectual as they do not consider different stages of platelet activation or the molecular dynamics involved in platelet interactions and are incapable to consider inter-individual variability. Here we propose a stochastic platelet deposition model and an inferential scheme to estimate the biologically meaningful model parameters using approximate Bayesian computation with a summary statistic that maximally discriminates between different types of patients. Inferred parameters from data collected on healthy volunteers and different patient types help us to identify specific biological parameters and hence biological reasoning behind the dysfunction for each type of patients. This work opens up an unprecedented opportunity of personalized pathology test for CVD detection and medical treatment.

Author summary

Cardiovascular accidents often result from blood deficiencies, such as platelets dysfunction. Current diagnosis techniques to detect such dysfunctions are not sufficiently accurate and unable to determine which platelet properties are affected. We develop a novel approach to describe in-vitro platelets deposition patterns in terms of clinically meaningful patient specific bio-physical quantities that allow for personalized clinical diagnostics. This approach combines mathematical modeling, statistical inference techniques, machine learning and high performance computation to estimate the values of these clinically relevant platelet properties. We demonstrate our approach on three classes of donors, healthy volunteers, patients subject to dialysis and patients with chronic obstructive

pulmonary disease. We claim that our approach opens a paradigm shift for the treatment and diagnosis of cardiovascular diseases, leading to personalized medicine.

Introduction

Cardio/cerebrovascular diseases (CVD) were the first cause of mortality worldwide in 2015, causing 31% of deaths according to World Health Organization [1]. Pathology tests for detection of CVD rely on testing functionality of blood platelets, which play a key role in the occurrence of these cardio/cerebrovascular accidents in addition to complex process of blood coagulation, involving adhesion, aggregation on the vascular wall to stop a hemorrhage while avoiding the vessel occlusion. In a comprehensive bio-medical evaluation study [2], the correlation between the clinical biological measures using platelet function tests and the occurrence of a cardiovascular event was found to be null for half of the techniques and rather modest for others, indicating the evident need for a more efficient tool or method to monitor patient platelet functionalities. The inadequacy of these tests can be explained by the fact that no current test allows for the analysis of the different stages of platelet activation or the prediction of the in-vivo behavior of those platelets [3, 4]. In addition, the current clinical tests do not take into account the dynamic aspect of the process of platelet aggregation and the role that red blood cells can have in this process. To address these issues, we extend here the stochastic model proposed in [5] that simulates numerically the deposition pattern of platelets observed in the Impact-R device [6], namely the sizes and number of aggregates as a function of time for a layer of whole blood subject to a controlled shear rate. This model is characterised by biophysically meaningful parameters the adhesion rate p_{Ad}, the aggregation rates p_{Ag} (of depositing on an already deposited platelet) and p_T (of depositing on an existing cluster of platelets), the deposition rate of albumin p_F, the attenuation factor a_T, and the flux of activated (AP) and non-activated platelets (NAP) obtained from their characteristics velocities (v_z^{AP} and v_z^{NAP}). The value of these parameters can be inferred by matching the simulation output with the corresponding in-vitro deposition pattern.

Our ***main claim*** here is that the values of some of these model parameters (eg. adhesion and aggregation rates) are precisely the information needed to assess various possible pathological conditions and to quantify their severity with reference to CVD.

To support this claim, we develop a methodology to identify medically interpretable parameters differentiating between patients and healthy volunteers. To infer the estimates of the biologically interpretable parameters of the stochastic platelet deposition model from the deposition patterns observed in the Impact-R device of platelet collected for a patient, we use approximate Bayesian computation (ABC) [7, 8] and report the estimated mode of the approximate posterior learned by ABC as the estimates, which can also be interpreted as approximate maximum likelihood estimates (MLE) when non-informative uniform priors are considered on the parameters. We note that ABC inferential algorithms depend on the choice of the summary statistics extracted from the datasets [7], which can be inferred via metric learning [9, 10] —a methodology that can provide the summary statistics able to maximally discriminate different patient groups. Leveraging on this crucial link between ABC and metric learning [11], we are able to identify medically meaningful parameters, which can distinguish between different types of patients and, at the same time, to estimate those parameters for each patient with the aim of developing a test for the pathology. We further notice that the proposed approach can be applied on each patient, in a systematic way. This reduces the bias of a human operator.

Finally, to verify our proposed methodology, we perform a four stage experiment: **1)** Collect blood or platelet from 32 patients (16 patients needing dialysis and 16 patients with Chronic Obstructive Pulmonary Disease—COPD) and 16 healthy volunteers; **2)** Study the deposition patterns observed in the Impact-R of platelet collected for each patient; **3)** Learn the summary statistics from this dataset which is able to maximally distinguish between the 3 types of patients; **4)** Estimate the model parameters for each of the patients, using ABC with the use of the learned discriminatory summary statistics from Stage 3. Studying the inferred parameters from each of the patients, we were able to identify medically meaningful parameters which are able to distinguish between patients of different types.

This study relies only on a small data set and is meant as a proof of concept. A forthcoming clinical study will provide a much larger data set on which we plan to demonstrate the potential of our approach convincingly.

Results and discussion

Dataset

The collected dataset (characterizing the deposition pattern in the Impact-R) can be divided into three groups: healthy volunteers (Group 1), patients needing dialysis (Group 2) and patients affected by Chronic Obstructive Pulmonary Disease (COPD) (Group 3). We collected blood samples from 16 volunteers or patients from each of the three groups. Half of the 16 patients having dialysis were also affected by diabetes and all the volunteers and patients were chosen from a broad age group.

Summary statistics learning and inference of parameters

Using Large Margin Nearest Neighbor Metric Learning (LMNN) [10], we first learn a 2-dimensional projection of the collected dataset, which maximally discriminates between different types of patients. We notice that the two features defining this two-dimensional space do not have a meaningful biological representation as they are just weighted non-linear combinations of the observed time-series of the platelet deposition pattern. We use these two features as summary statistics for ABC and thus the fact that they have no biological interpretation is not relevant since they are only used to facilitate the estimation process of biologically meaningful model parameters that are, in turn, used to define the clinical tests. Next we use the Euclidean distance on this projection space for ABC to infer the parameters of the stochastic platelets depositions model for each of the volunteers (or patients), using the corresponding deposition and aggregation pattern of the platelets in their blood displayed in Impact-R machine. This provides us with a posterior distribution of the parameters given the data from each individual volunteer (patient). After learning the posterior distribution, to provide an estimate of the parameters for each of the patients, we calculate the maximum a posteriori (MAP) estimate of the parameters. Here we note that the Bayesian point estimates are minimizers of posterior loss (eg. posterior mean minimizing squared error loss or MAP minimizing 0–1 loss) [12]. These ABC inferred parameters provide valuable biological interpretation. In Fig 1, we illustrate the inferred posterior distribution for a patient with COPD and the corresponding maximum a posteriori (MAP) estimate of the parameters.

Uncertainty in MAP estimation

We illustrate the uncertainty of the MAP estimates in each patient group through the boxplots of MAP estimates for each patient in each of the three groups in Fig 2 for the three most
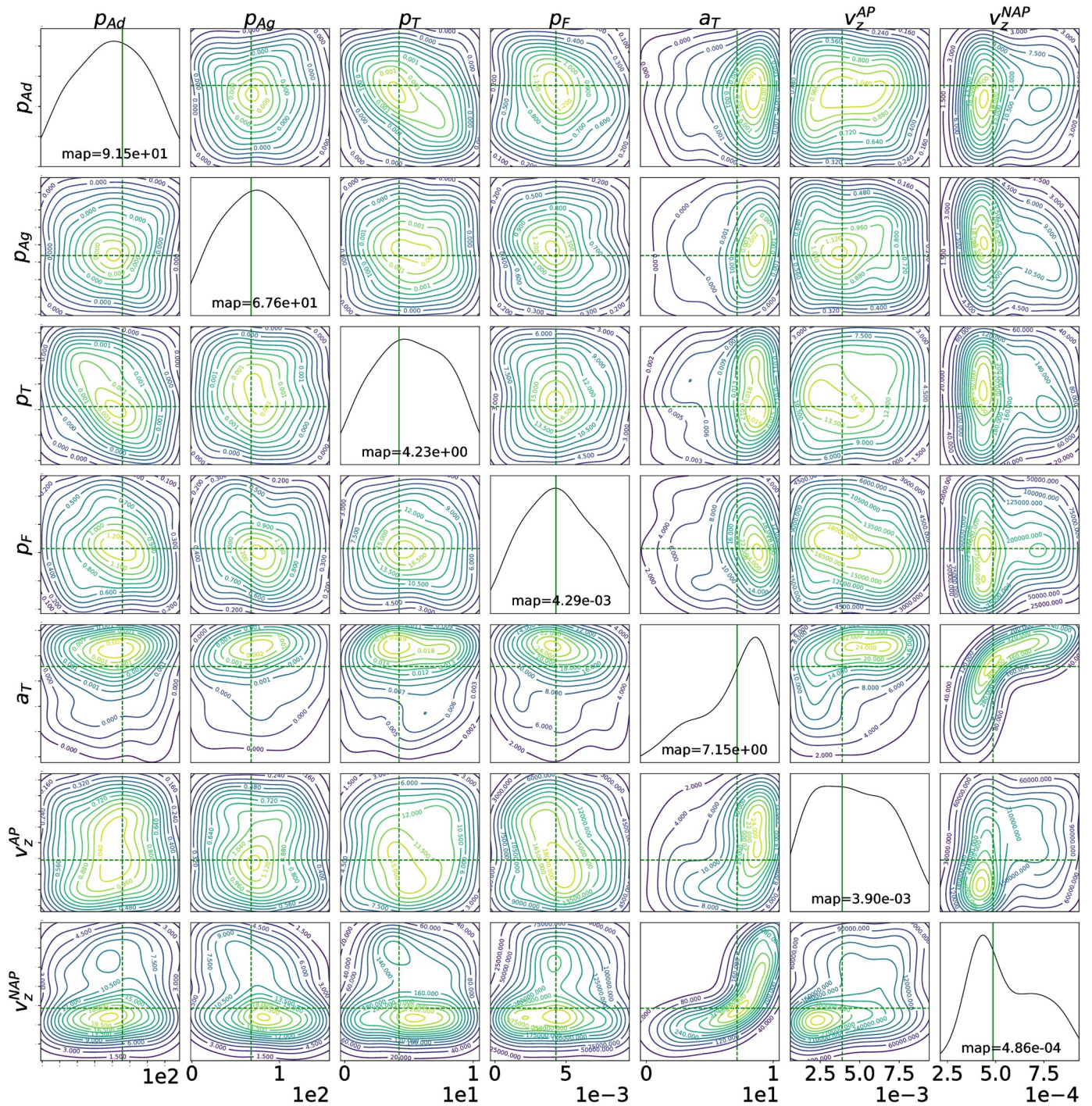
Fig 1. Posterior distribution of the model parameters for a COPD patient. The green cross indicates the maximum a posteriori (MAP) estimate of the parameters p_{Ad} $[s^{-1}]$, p_{Ag} $[s^{-1}]$, p_T $[s^{-1}]$, p_F $[s^{-1}]$, a_T $[\mu m^2 s^{-1}]$, v_z^{AP} $[ms^{-1}]$, v_z^{NAP} $[ms^{-1}]$.

discriminative parameters. The average observed standard deviation across the three groups are $\sigma^*_{p_{Ag}} : 17.79, \sigma^*_{a_T} : 1.64, \sigma^*_{v_z^{NAP}} : 1.6e - 04$.

One of our main assumption to build the test for pathology is that the parameter values for each of the patients in a specific group are centered around a true value of the parameters with

Fig 2. Boxplot of the MAP estimates of the biologically meaningful discriminating parameters grouped according to patient types.
Healthy: Healthy volunteers (16 patients); Dialysis: All Patients undergoing dialysis (16 patients); COPD: Patients with COPD (16 patients). The box in the plot extends from the first quartile (Q1) to third quartile (Q3) of the estimated parameters for each type of patients, with an orange horizontal line at the median. The upper whisker extends to the largest value less than Q3 + 1.5 * IQR and the lower whisker extends t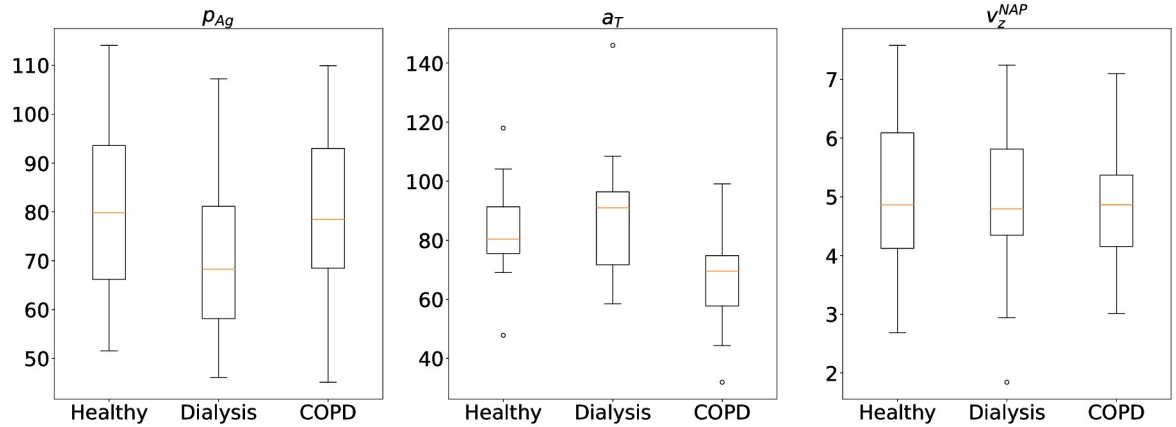o the lowest value greater than Q1 − 1.5 * IQR, where IQR is the interquartile range (Q3-Q1). Beyond the whiskers, any values are considered outliers and are plotted as individual points.

a small variance. Under this assumption, using consistency theorem for MAP estimates [13] we can argue that the MAP estimates for each patient in that group would converge to the true value if we have increasingly many samples (eg. repeated measurements using Impact-R machine) for each of the patients. Given we only have one sample from each of the patient, we can argue that the distribution of the MAP estimates of the patients in a group would be centered around the true value. To justify the standard deviation observed in the MAP estimates of the patients in a group, we simulated 10 dataset using a true parameter value and computed the standard deviation of the MAP estimates for each of the simulated dataset ($\hat{\sigma}_{p_{Ag}} : 23.4, \hat{\sigma}_{a_T} : 0.95, \hat{\sigma}_{v_z^{NAP}} : 2.71e - 05$). We notice that the variance in the MAP estimates of the real data in a group [Fig 2] is similar, in its order of magnitude, to the one observed in our simulation study.

Finally, we notice that the mean/median of the MAP estimates in each group of patients are significantly different among the three groups for the parameters p_{Ag}, a_T, v_z^{NAP}, possibly indicating a discriminative behaviour among the three groups, which helped us to identify the most discriminative parameters and devise a test for pathology.

Kruskal-Wallis H-test

Our main claim is based on the assumption that the median of the MAP estimates of the three patient groups are different among the groups for some of the parameters. We notice that the distribution of the MAP estimates of patients from different groups can still overlap due to the uncertainty in MAP estimation. To test whether the patient specific estimated parameters of the model can distinguish patients from healthy volunteers and furthermore discriminate between different types of the patients, we use the Kruskal-Wallis H-test [15] which tests *the null hypothesis that the median of the MAP estimates of the different types of patients are equal versus a bilateral alternative*. The computed test statistics and P-values between all the three groups (healthy volunteers, patients needing dialysis and patients with COPD) are reported in the first column of Table 1. Considering a cutoff for significance of 0.05, the null hypothesis gets rejected for the parameters p_{Ag} and v_z^{NAP} indicating that the values of these parameters significantly differ between the groups.

Table 1. Statistics (P-values, P-values corrected for multiple testing) of Kruskal-Wallis test using maximum a posteriori (MAP) estimates of the parameters. A higher value of the statistics and P-values smaller than 0.05 (bold) indicates that the medians of the estimated parameters of the corresponding groups are statistically different with a significance cutoff of 0.05. Healthy indicates healthy volunteers, Dialysis indicates patients under dialysis, and COPD stands for patients with COPD. The correction of P-values for multiple testing has been done using the Benjamini-Hochberg procedure [14].

Parameters	All three classes	Healthy vs Dialysis	Healthy vs COPD	Dialysis vs COPD
p_{Ad}	2.7 (2.5e-1, 3.5e-1)	2.2 (1.3e-1, 2.3e-1)	3.5e-2 (8.5e-1, 9.1e-1)	1.7 (1.8e-1, 2.4e-1)
p_{Ag}	**9.3 (9.5e-3, 4.9e-2)**	2.4e-1 (6.2e-1, 7.9e-1)	**7.3 (6.6e-3, 4.6e-2)**	**6.3 (1.1e-2, 4e-2)**
p_T	2.3e-2 (9.8e-1, 9.8e-1)	1.4e-3 (9.7e-1, 9.7e-1)	2.2e-2 (8.8e-1, 9.1e-1)	1.3e-2 (9.1e-1, 9.1e-1)
p_F	1.4 (4.9e-1, 5.7e-1)	1.7e-1 (6.7e-1, 7.9e-1)	4.1e-1 (5.2e-1, 9.1e-1)	1.5 (2.1e-1, 2.4e-1)
a_T	4.9 (8.2e-2, 1.9e-1)	**3.8 (5.0e-2, 2.2e-1)**	1.2e-2 (9.1e-1, 9.1e-1)	3.5 (5.9e-2, 1.3e-1)
v_z^{AP}	3.7 (1.5e-1, 2.7e-1)	2.7 (9.7e-2, 2.2e-1)	3.5e-2 (8.5e-1, 9.1e-1)	2.7 (9.7e-2, 1.7e-1)
v_z^{NAP}	**8.5 (1.4e-2, 4.9e-2)**	2.8 (8.9e-2, 2.2e-1)	2.5 (1.1e-1, 3.9e-1)	**7.3 (6.6e-3, 4e-2)**

https://doi.org/10.1371/journal.pcbi.1009910.t001

We notice that the rejection of the null hypothesis does not indicate which of these groups differ. Hence, we perform post-hoc Kruskal-Wallis H-test between healthy volunteers and patients needing dialysis, healthy volunteers and patients with COPD and between patients needing dialysis and patients with COPD [columns 2, 3 and 4 of Table 1 respectively]. This indicates that p_{Ag} can clearly differentiate patients with COPD from both healthy volunteers and patients having dialysis. Further a_T and v_z^{NAP} being able to discriminate patients needing dialysis correspondingly from healthy volunteers and patients with COPD. We list these parameters which are capable to distinguish between the corresponding groups in Table 2.

Discriminating parameters and pathology test

According to our analysis, the biologically meaningful parameters which are able to discriminate between different patient types up to some accuracy are p_{Ag}, a_T and v_z^{NAP}, as shown in Table 2 and Fig 2. Further, these parameters can be divided into two distinct types of pathologies related to biochemical and conformational changes correspondingly in platelets and red blood cells (RBCs).

Pathological changes in platelets. The first group of parameters (p_{Ag}, a_T) represent explicit intrinsic changes in platelets associated with the presence of pathology, which causes a change in their patterns of adhesion and aggregation. The common thread between dialysis patients and COPD patients is the existence of chronic systemic inflammation implicated in the development of cardiovascular disease. In response to inflammation, it is well known that platelets in COPD and dialysis patients are activated in the bloodstream, altering their hemostatic properties [16–19] and therefore the process of adhesion and aggregation.

Biochemical and conformational changes in RBCs. v_z^{NAP} reflects another aspect of the presence of pathology, where changes in the velocity of platelets are caused via their interaction with red blood cells (RBCs), which may have undergone biochemical and conformational changes altering blood rheology [20] under pathological situations. These changes in RBCs are known as spherization and have been observed in sepsis, dialysis patients, COPD patients and

Table 2. Biologically meaningful discriminating parameters. The parameters which are significantly different between the corresponding two groups of patients (volunteers).

	Discriminating parameters
Healthy vs Dialysis	a_T
Healthy vs COPD	p_{Ag}
Dialysis vs COPD	p_{Ag}, v_z^{NAP}

https://doi.org/10.1371/journal.pcbi.1009910.t002

Table 3. Sensitivity and specificity of the proposed test using the most discriminative parameters to identify diseased patients compared to healthy volunteers.

	Healthy vs COPD	Healthy vs Dialysis
Discriminating parameter	p_{Ag}	a_T
Sensitivity	0.75	0.56
Specificity	0.75	0.62

other pathologies causing chronic or acute systemic inflammation [21]. Recently, [22] reported that the RBC spherization induces an increase in platelet adhesion and aggregation processes and an increase in platelet transport to the wall. Hence, the changes observed in platelets velocities under pathologies may indicate medical conditions which causes biochemical and conformational changes to RBCs.

Pathology test. Based on the most discriminative parameters (a_T and p_{Ag}) identified by our analysis, we devise a test for pathology to identify diseased patients (correspondingly for patients with COPD and patients having dialysis). An individual is identified as healthy or belonging to COPD group (similarly to dialysis patients) if the MAP estimate of the parameter a_T (correspondingly p_{Ag}) is closer to the median of the MAP estimates of the healthy volunteers or closer to the corresponding value estimated on patients in the COPD group (respectively dialysis patients). The sensitivity and specificity [23] of our two tests are reported in the Table 3. This shows that we can identify patients having COPD with higher degree of accuracy given that our analysis only depends on a relatively small number of patients/volunteers.

Materials and methods

Ethics statement

Volunteers were recruited at the nephrology and pneumology units of the CHU-Charleroi, ISPPC Hôpital Vésale in Belgium. Written informed consent was obtained from each patient and healthy donor included in the study. The protocol of the study was in conformity with the ethical guidelines of the Helsinki Declaration of 1975 (revised in 2000) and was approved by the institution's ethics committee (No: OM008; P17/49_27/09). Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Based on detailed *in vitro* experiments using the Impact-R device mimicking platelet adhesion-aggregation in blood vessels, first we provide a model which is an *in silico* counterpart for an in-depth description and understanding of the phenomenon and the underlying mechanisms.

Impact-R experiment

Impact-R [6], a well-known platelet function analyzer, is a cylindrical device whose bottom wall is a fixed disc (deposition substrate), while the upper wall is a rotating disc (shaped as a cone with a small angle). The height of the device is 0.82 mm and due to the motion of the upper wall a pure shear flow is created. A controlled shear rate $\dot{\gamma}$ is produced in a given observation window of $1 \times 1 \ mm^2$, where we track the formation of clusters resulting from the deposition and aggregation of platelets. Blood was drawn from both healthy and diseased donors with different hematocrit (volume fraction of red blood cells (RBC)). Before starting the tests, a sample is recovered and analysed, to determine the concentration of activated (AP) and non-activated (NAP) platelets. Serum albumin, the most abundant protein in human blood plasma, antagonises with the platelets, preventing them from adhering to the substrate. The quantities
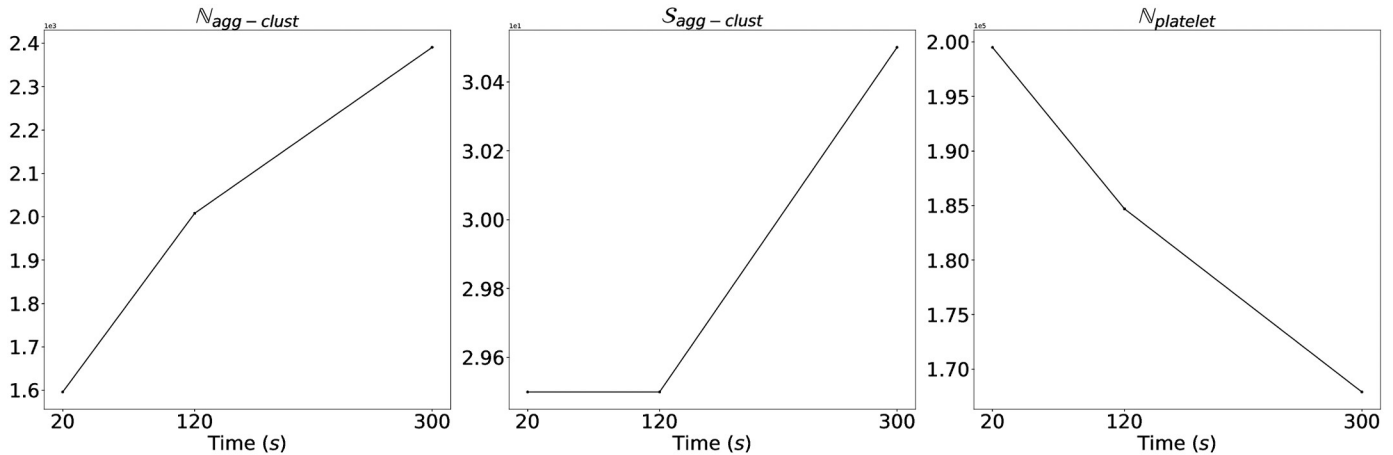
Fig 3. Data from Impact-R device. The data collected from Impact-R for a healthy volunteer containing three observed quantities $\mathbb{N}_{agg-clust}(t)$, $\mathcal{S}_{agg-clust}(t)$ and $\mathbb{N}_{platelet}(t)$ are correspondingly average size of the aggregation clusters, their number per mm^2 and the number of non-activated platelets per $\mu\ell$ still in suspension at time t.

of interest are the number of clusters and their size formed in the substrate, and the number of AP still in suspension. Our goal is to explain the observed (*in vitro* experiments) time evolution of these three quantities, which is illustrated for a healthy volunteer in Fig 3.

Stochastic model of platelet deposition

The deposition process in Impact-R was successfully described with a mathematical model for the first time in [24, 25] accounting for the following observations: (i) AP adhere to the deposition surface, forming a seed for a new cluster, (ii) NAP and AP can deposit at the periphery or on top of an existing cluster and (iii) Albumin (Al) deposits on the surface, thus reducing locally the adhesion and aggregation rates of platelets. A sketch of the situation shown in Fig 4. In [25], the platelets reach the bottom layer due to a RBC-enhanced shear-induced 1D diffusion. In the present work, we propose a fully stochastic model of platelet deposition, by substituting the 1D diffusion systems with a 3D random walk, while keeping the deposition dynamics the same. The reason of this new approach is to avoid the introduction of a boundary layer (denoted by Δz in [25]) when coupling platelet transport with platelet deposition. With this particle based approach, a platelet becomes a candidate for deposition whenever it hits the deposition surface. In [25] the diffusion coefficient D and the thickness of the boundary layer Δz where determined independently of the other model parameters, precisely by assuming a random motion of the platelets. Now, we keep the same level of description everywhere and, instead of D and Δz, two new parameters are considered, namely the characteristic velocities of activated and non-activated platelets. They will be inferred from the data, together with the adhesion and aggregation rates, and other quantities defined below.

The random walk of the platelets can be described by their jump $(\Delta z(t), \Delta x(t), \Delta y(t))$ at each iteration t:

$$\Delta z(t) = \lambda \ v_z \ |s_z| \ dt, \tag{1}$$

$$\Delta x(t) = v_{xy} \ |s_{xy}| \ \cos(2\pi r) \ dt, \tag{2}$$

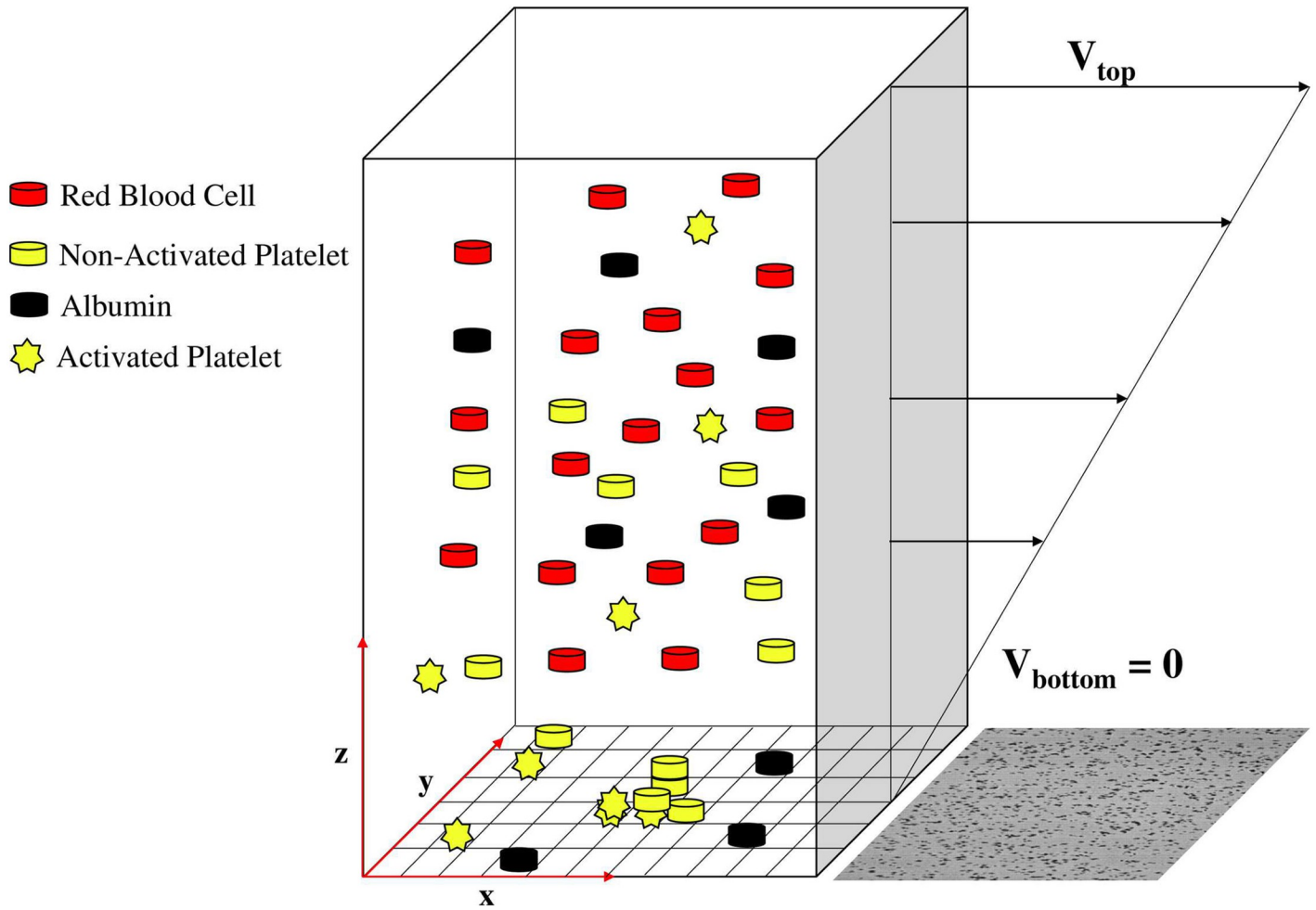$$\Delta y(t) = v_{xy} \ |s_{xy}| \ \sin(2\pi r) \ dt, \tag{3}$$

Fig 4. Window of Impact-R device. The bottom wall is a fixed boundary of dimensions $1 \times 1\ mm^2$, the wall-bounded direction is 0.82 mm. The bulk contains whole blood at different hematocrit. The discretization of the substrate is such that in every cell can fit just one platelet. The initial densities of the blood particles are determined by the *in vitro* experiment and usually are about: 172 200 $(\mu l)^{-1}$ for NAP, 4808 $(\mu l)^{-1}$ for AP, $2.69 \times 10^{13}\ (\mu l)^{-1}$ for Al. The image next to the discretised substrate corresponds to the *in vitro* experiment.

https://doi.org/10.1371/journal.pcbi.1009910.g004

where $v_{z,xy}$ is a speed unit, r is a random variable uniformly distributed in $[0, 1]$, $s_{z,xy} \in (-\infty, \infty)$ is a random variable distributed as a standard normal distribution, $\lambda \in \{-1, 1\}$ with probability 1/2 for each outcome, and dt is the time step of the simulation. Superimposing the stochastic motion with the velocity field of the pure shear flow, the positions of the platelets are updated as

$$z_i(t + dt) = z_i(t) + \Delta z_i(t), \tag{4}$$

$$x_i(t + dt) = x_i(t) + \Delta x_i(t) + \dot{\gamma}\, z_i\, dt, \tag{5}$$

$$y_i(t + dt) = y_i(t) + \Delta y_i(t). \tag{6}$$

It should be noted that this stochastic model is physically equivalent to the shear-induced diffusion model used in [25]. From the random motion of particles, diffusion constants emerge both along the flow direction and perpendicular to it. The equivalence of random walk

and diffusion is a well known result, see for instance [26]. It has been checked that the platelet deposition pattern obtained with this new model is the same, up to statistical fluctuations, as with the model used in [25], when the physical properties are the same. The present model is computationally more costly than that of [25], but provides more flexibility for parameter inference, in particular to infer, from in vitro data, the transport properties of platelets towards the deposition surface. Additionally the present transport model no longer assumes the same transport properties for activated and non-activated platelets. Actually, this feature allows us to better explain the Impact-R data as activated platelets turn out to move faster, probably due to their increased effective hydrodynamic size. As a consequence, the final results of both models cannot be compared.

Our interest here is to model the transport of platelets in the direction perpendicular to the blood flow so as to obtain the flux of platelets that will reach the so-called Cell Free Layer (the layer without red blood cells near a wall), at the bottom of the Impact-R device and be candidate for deposition. The stochastic transport along the blood flow direction is not expected to have an impact on the deposition, due to the fast mixing of platelet in the horizontal plan. This assumption made in [25] is confirmed by the present model which includes explicitly this horizontal transport. Note finally that here we can exclude any drift in the vertical direction due to the up-down symmetry in the Impact-R setting. This absence of a drift, sometimes proposed to describe platelet transport in a tube, has been confirmed by full resolved blood flow simulations, in which deformable red blood cells and platelets are in suspension in a plasma subject to a shear flow [27].

Owing to the different dynamics and physics governing the activated and non-activated platelets, they have in principle different speed units ($v_{z,xy}^{AP}$, $v_{z,xy}^{NAP}$). Regarding the motion of albumin, its abundance allows us to neglect the small density gradients due to its deposition, and thus albumin can deposit at any time at a maximum value of deposition rate.

The AP and NAP that cross the lower boundary of the computational domain are removed from the bulk if the deposition is successful, or get trapped at the Cell Free Layer (CFL) for a future deposition attempt (they never get re-injected into the bulk). Further, periodic conditions are applied at the x, y directions, and bounce back boundary condition for the platelets that cross the upper boundary. Assuming a good horizontal mixing in the xy-plane due to the rotating flow, and given its low impact on the deposition process, the stochastic part of the motion in the x, y directions can be fixed to the same order of magnitude as the z direction velocity. Therefore, we consider $v_{xy}^{AP} = v_z^{AP}$, $v_{xy}^{NAP} = v_z^{NAP}$ and $s_{xy} = s_z$.

Next we describe the deposition rates. Let us denote by $N_{i,j}(t)$ the number of candidate particles for deposition above the cell at position i, j of the discretised substrate. The deposition of the platelets and albumin on the substrate follow the stochastic rules described in [25], i.e., based on the $N_{i,j}(t)$ and on the occupancy of the i, j-th cell at time t. Albumin that reaches the substrate at time t deposits with a probability $P(t)$ which depends on the local density $\rho_{al}(t)$ of already deposited albumin. We assume that P is proportional to the remaining free space in the cell,

$$P(t) = p_F(\rho_{max} - \rho_{al}(t))dt, \tag{7}$$

where p_F is a parameter to be determined and ρ_{max} is given by the constraint that at most 100,000 albumin particles can fit in a deposition cell of area $\Delta S = 5 \ (\mu m)^2$, corresponding to the size of a deposited platelet (obtained as the smallest variation of cluster area observed with the microscope). An activated platelet that hits a platelet-free cell deposits with a probability Q, where Q decreases as the local concentration ρ_{al} of albumin increases. We assumed that

$$Q = p_{Ad} \exp(-a_T \rho_{al})dt, \tag{8}$$

where p_{Ad} and a_T are parameters to be determined. This expression can be justified by the fact that a platelet needs more free space than an albumin to attach to the substrate, due to their size difference. In other words, the probability of having enough space for a platelet, decreases roughly exponentially with the density of albumin in the substrate (more details in [5]). In our model, AP and NAP can deposit next to already deposited platelets. From the above discussion, the aggregation probability R is assumed to be

$$R = p_{Ag} \exp(-a_T \rho_{al})dt, \tag{9}$$

with p_{Ag} another unknown parameter. We also introduce p_T the rate at which platelets deposit on top of an existing cluster. Fig 4 presents coarsely the competing adhesion-aggregation process between albumin and platelets. More details on the stochastic deposition rules can be found in [25].

For the purpose of the present study, the platelet deposition model $\mathcal{M}$ is parametrized in terms of the seven quantities introduced above, namely the adhesion rate p_{Ad}, the aggregation rates p_{Ag} and p_T, the deposition rate of albumin p_F, the attenuation factor a_T, and the velocities of AP and NAP v_z^{AP} and v_z^{NAP}. Collectively, we define

$$\boldsymbol{\theta} = (p_{Ag}, p_{Ad}, p_T, p_F, a_T, v_z^{AP}, v_z^{NAP}).$$

If the initial number of AP and NAP at time $t = 0$ ($\mathbb{N}_{platelet}(0)$ and $\mathbb{N}_{act-platelet}(0)$), as well as the concentration of albumin are known from the experiment, we can forward simulate the deposition of platelets over time using model $\mathcal{M}$ for the given values of these parameters $\boldsymbol{\theta} = \boldsymbol{\theta}^*$:

$$\mathcal{M}[\boldsymbol{\theta} = \boldsymbol{\theta}^*] \rightarrow \{(\mathbb{N}_{agg-clust}(t), \mathcal{S}_{agg-clust}(t), \mathbb{N}_{platelet}(t)), \ t = 0, \dots, T\}.$$

where $\mathbb{N}_{agg-clust}(t)$,$\mathcal{S}_{agg-clust}(t)$ and $\mathbb{N}_{platelet}(t)$ are correspondingly average size of the aggregation clusters, their number per mm^2, the number of non-activated and pre-activated platelets per $\mu\ell$ still in suspension at time t.

The Impact-R experiments have been repeated with the whole blood obtained from each of the volunteers and patients and the observations were made at time, 20 sec., 120 sec. and 300 sec. At these three time points, $(\mathbb{N}_{agg-clust}(t), \mathcal{S}_{agg-clust}(t), \mathbb{N}_{platelet}(t))$ are measured [Fig 3]. Let us call the observed dataset collected through experiment as,

$$\boldsymbol{x^0} \equiv \{(\mathbb{N}^0_{agg-clust}(t), \mathcal{S}^0_{agg-clust}(t), \mathbb{N}^0_{platelet}(t)) : t = 0 \ \text{sec.}, \dots, 300 \ \text{sec.}\}.$$

Estimation of model parameters

As the likelihood function induced by the platelets deposition model is analytically intractable due to the need of computing a very high-dimensional integral, we can not compute the maximum likelihood estimate of the model parameters or perform traditional Bayesian inference. In setting where the likelihood function is not available, approximate Bayesian computation (ABC) [7] offers a way to sample from an approximate posterior distribution of the parameter $p(\boldsymbol{\theta}|\boldsymbol{x^0}) \approx \pi(\boldsymbol{\theta})p(\boldsymbol{x^0}|\boldsymbol{\theta})$ given the observed data $\boldsymbol{x^0}$, where $\pi(\boldsymbol{\theta})$ and $p(\boldsymbol{x^0}|\boldsymbol{\theta})$ are correspondingly the prior distribution on the parameter $\boldsymbol{\theta}$ and the likelihood function. Further we note that the mode of this approximate posterior distribution (i.e. the maximum-a-posteriori -MAP- estimate) is also the approximate maximum likelihood estimate if we assume the prior distribution on the parameter to be uniform. Following this, given the observed data, we compute the MAP estimates using the ABC approximate posteriors of the parameters of the stochastic platelet deposition model.
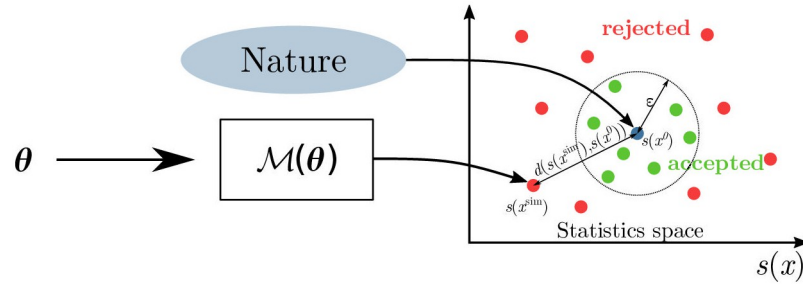
Fig 5. Approximate Bayesian computation. Having observed data $\boldsymbol{x^0}$ from an individual patient (the gray dot), we sample parameter values θ from the prior and generate observations through the model simulator, $\mathcal{M}(\boldsymbol{\theta})$, that are then accepted (green) or rejected (red) according to their distance from the observation measured by $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|s(\boldsymbol{x}_1) - s(\boldsymbol{x}_2))\|_2$ on the summary statistics space.

Approximate Bayesian computation (ABC)

The fundamental ABC rejection sampling scheme iterates the following steps:

1. Draw $\boldsymbol{\theta}$ from the prior $\pi(\boldsymbol{\theta})$.

2. Simulate a synthetic dataset $\boldsymbol{x^{\text{sim}}}$ from the simulator-based model $\mathcal{M}(\boldsymbol{\theta})$.

3. Accept the parameter value $\boldsymbol{\theta}$ if $d(\boldsymbol{x^{\text{sim}}}, \boldsymbol{x^0}) < \epsilon$. Otherwise, reject $\boldsymbol{\theta}$.

See Fig 5 for a visualization of the above algorithm.

Here, the metric on the dataspace $d(\boldsymbol{x^{\text{sim}}}, \boldsymbol{x^0})$ measures the closeness between $\boldsymbol{x^{\text{sim}}}$ and $\boldsymbol{x^0}$. The accepted $(\boldsymbol{\theta}, \boldsymbol{x^{\text{sim}}})$ pairs are thus jointly sampled from a distribution proportional to $\pi(\boldsymbol{\theta})$ $p_{d,\boldsymbol{\epsilon}}(\boldsymbol{x^0}|\boldsymbol{\theta})$, where $p_{d,\boldsymbol{\epsilon}}(\boldsymbol{x^0}|\boldsymbol{\theta})$ is an approximation to the likelihood function $p(\boldsymbol{x^0}|\boldsymbol{\theta})$:

$$p_{d,\boldsymbol{\epsilon}}(\boldsymbol{x^0}|\boldsymbol{\theta}) = \int p(\boldsymbol{x^{\text{sim}}}|\boldsymbol{\theta})\mathbb{K}_{\boldsymbol{\epsilon}}(d(\boldsymbol{x^{\text{sim}}}, \boldsymbol{x^0}))d\boldsymbol{x^{\text{sim}}}, \qquad (10)$$

where $\mathbb{K}_{\boldsymbol{\epsilon}}(d(\boldsymbol{x^{\text{sim}}}, \boldsymbol{x^0}))$ is in this case a probability density function proportional to $\mathbb{1}(d(\boldsymbol{x^{\text{sim}}}, \boldsymbol{x^0}) < \boldsymbol{\epsilon})$ and $\mathbb{1}(\cdot)$ is used as an indicator function. Besides this choice for $\mathbb{K}_{\boldsymbol{\epsilon}}(d(\boldsymbol{x^{\text{sim}}}, \boldsymbol{x^0}))$, that has been exploited in several papers (for instance [28–31]), ABC algorithms relying on other choices exist, for instance with the kernel $\mathbb{K}$ being $\exp(-d(\boldsymbol{x^{\text{sim}}}, \boldsymbol{x^0})/\boldsymbol{\epsilon})$ as in simulated-annealing ABC (SABC) [32]. More advanced algorithms than the simple rejection scheme detailed above are possible, for instance ones based on Sequential Monte Carlo [30, 31], in which various parameter-data pairs are considered at a time and are evolved over several generations, while $\boldsymbol{\epsilon}$ is decreased towards 0 at each generation to improve the approximation of the likelihood function, so that we are able to approximately sample from the true posterior distribution. For inference of parameters of the platelets deposition model, here we choose the SABC algorithm, based on its suitability to high performance computing systems [33]. For practical implementation, SABC was run for 20 iterations generating 510 samples from the posterior distribution of the model parameters given data from each patient, keeping all other parameters fixed to the default values proposed in the Python package 'ABCpy' [34]. Next, we explain how the distance between datasets used for ABC was chosen via a discriminative summary statistics learning approach and finally how the MAP estimates (MAP) of the parameters was computed.

Discriminative summary statistics learning (DSSL)

Traditionally, distance between $\boldsymbol{x^{\text{sim}}}$ and $\boldsymbol{x^0}$ is defined by summing over Euclidean distances between all possible pairs composed by one simulated and one observed datapoint in the

corresponding datasets. Recently, distances for ABC have also been defined through classification accuracy [35], Kullback-Liebler divergence [36], maximum mean discrepancy [37] or by Wasserstein distance [38], under the assumption that the datapoints in each datasets are identical and independently distributed and they are available in a large number in both $x^{\mathbf{sim}}$ and $x^{\mathbf{0}}$. We notice that in our setting we only have one datapoint in the observed dataset corresponding to the platelet deposition pattern of a patient and also due to the very expensive nature of our simulator model (simulation of one datapoint takes around 15 minutes) we can only have few datapoints in the simulated dataset. Hence, here we concentrate on the definition of distances through Euclidean distance on summary statistics extracted from the dataset when we only have one data-point in both $x^{\mathbf{sim}}$ and $x^{\mathbf{0}}$.

When the data x is high-dimensional, a common practice in ABC literature is to define d as Euclidean distance between a lower-dimensional summary statistics $s : x^{\mathbf{sim}} \mapsto s(x^{\mathbf{sim}})$. Reducing the data to suitably chosen summary statistics may also yield more robust inference with respect to noise in the data. Moreover, if the statistics is sufficient, then the above modification provides us with a consistent posterior approximation [39], meaning that we are still guaranteed to converge to the true posterior distribution in the limit $\epsilon \to 0$. As sufficient summary statistics are not known for the majority of the complex models, the choice of summary statistics remains a problem [40] and they have been previously constructed using neural networks trained on a 'pilot' simulated dataset from the simulator model as in neural network based semi automatic summary statistics learning (SASL) [41, 42] or summary statistics learning minimizing triplet loss (TLSL) [11]. Detailed description of SASL and TLSL can be found in S1 Appendix.

These procedures are computationally expensive due to the need of the simulation of the 'pilot' dataset, further the learned summary statistics using these methods are not able to discriminate between datasets from different patient types. As the main goal of the present research is to learn parameter values which are able to differentiate between different patient types, here we propose a methodology to learn such summary statistics and name it as discriminative summary statistics learning (DSSL). Learning summary statistics which is most discriminative between datasets with different labels falls under a well-developed field of research in metric-learning [9]. We use Large Margin Nearest Neighbor Metric Learning (LMNN) [10], one of the metric-learning approaches, which learns a Mahalanobis distance between data from any two patients x_1 and x_2 able to discriminate between datasets with different labels,

$$d_M(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)} \tag{11}$$

where M is a $d \times d$ positive semi-definite matrix. We note that learning of the Mahalanobis distance here corresponds to learning a summary statistics of the data. It is sufficient to recall that for each positive semidefinite matrix M there exists a square matrix L such that $M = L^T L$. Therefore, we can write Eq 11 in the following way:

$$d_M(x_1, x_2) = \sqrt{(x_1 - x_2)^T L^T L (x_1 - x_2)} = \|L(x_1 - x_2)\|_2,$$

where $\|\cdot\|_2$ is the Euclidean distance, from which it is clear that the above corresponds to learning the transformation $s : x \mapsto s(x) = Lx$ which is able to discriminate between different patient groups.

To learn the transformation, LMNN solves the following optimization problem:

$$\min_L \sum_{i,j} \eta_{ij} \|L(x_i - x_j)\|^2 + \sum_{i,j,l} \eta_{ij}(1 - y_{ij})[1 + \|L(x_i - x_j)\|^2 + \|L(x_i - x_l)\|^2]_+,$$

where x_i is a data from a patient, x_j is one of that patient's k-nearest neighbors belonging to the same group of patients, and x_l are all the other data from patients of different type within the neighborhood, $\eta_{ij}, y_{ij} \in \{0, 1\}$ are indicators, $\eta_{ij} = 1$ if x_j is one of the k-nearest neighbors (conditioned on being of the same type of patient as x_i) of x_i, $y_{ij} = 0$ indicates x_i, x_j are different types of patients, $[\cdot]_+ = max(0, \cdot)$ is the Hinge loss. Intuitively, LMNN tries to learn a metric able to keeps k-nearest neighbors from the same patient group close together, while keeping patients from the other groups well separated. Further we note LMNN does not make any assumptions about the distribution of the data. To learn the projection by using LMNN, we consider the original data, its second and third order polynomial expansion and the cross products between them. Further, we manually tuned the tuning parameters for the LMNN algorithm provided in Python package 'metric-learn' [43] to maximize the rand index [44] between the true patient clusters in the dataset and the clustering achieved using agglomerative hierarchical clustering with the Euclidean distance on the learned discriminative summary statistics. Euclidean distance between the learned summary statistics from the data of different patient types were able to cluster the dataset with 100% accuracy. In Fig 6 we illustrate the learned discriminative summary statistics space in which the 3 groups of patients/volunteers are accurately clustered.

To validate our DSSL approach, we compare DSSL with SASL and TLSL by using posterior predictive checks for a experimental simulated dataset from the stochastic platelet deposition model. Further experimental details can be found in S1 Appendix. The main goal here is to
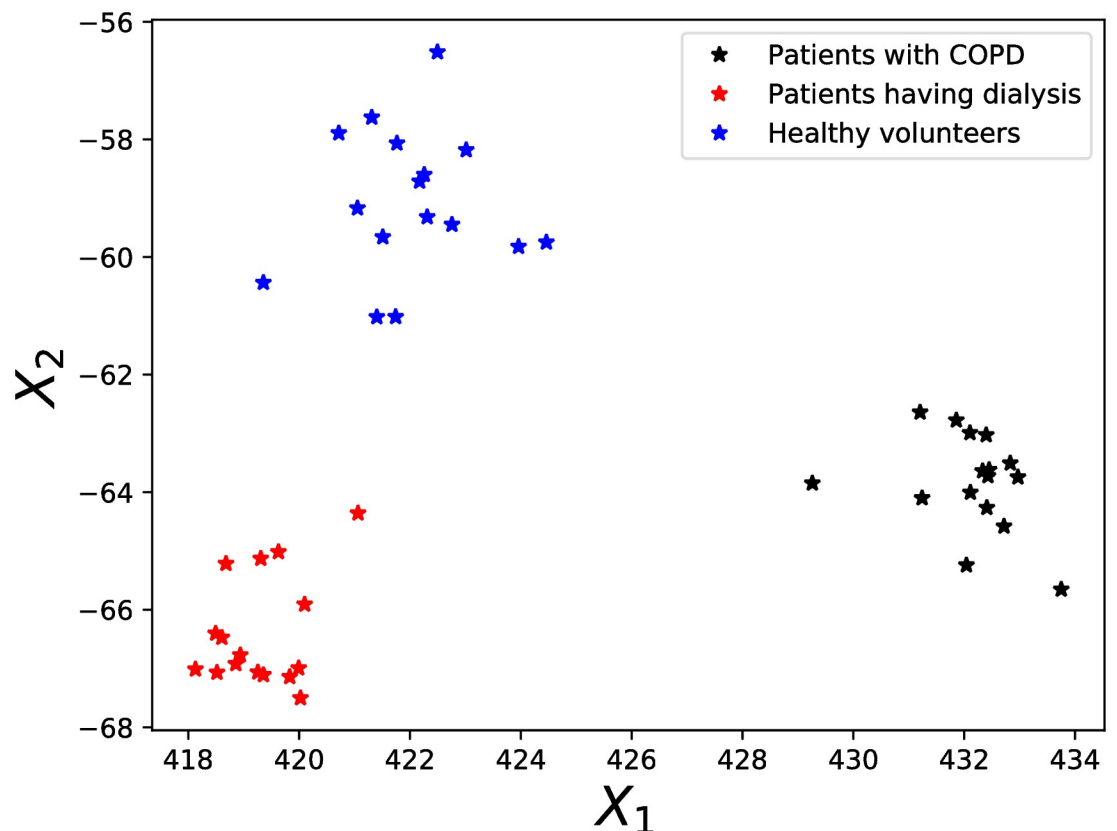

Fig 6. Discriminative summary statistics space. The discriminative summary statistics space learned by DSSL, in which the patients with COPD, patients having dialysis and healthy volunteers were accurately clustered.
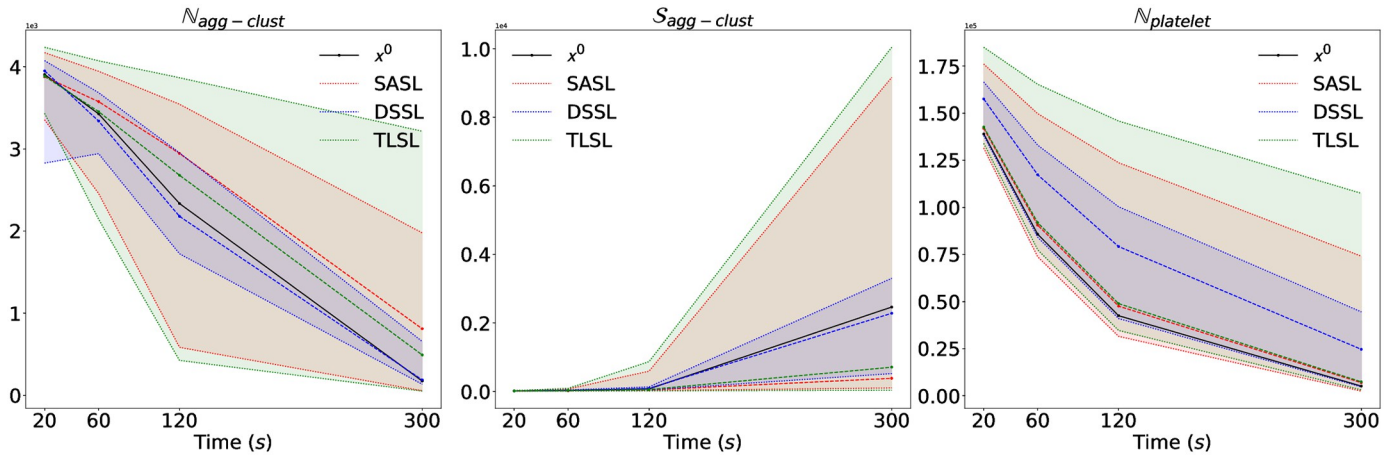
https://doi.org/10.1371/journal.pcbi.1009910.g006

Fig 7. Comparison of predictive performance of SASL, DSSL and TLSL. The experimental data (black solid line), 95% predictive credibility interval (colored shaded area) and the median prediction (colored dashed line) for SASL (red), DSSL (blue) and TLSL (green) are illustrated to compare the predictive performance of SASL, DSSL and TLSL summary learning approach for $\mathbb{N}_{agg-clust}$, $\mathcal{S}_{agg-clust}$ and $\mathbb{N}_{platelet}$.

analyze the degree to which the experimental data deviate from the data generated from the inferred posterior distribution of the parameters. Hence we want to generate data from the model using parameters drawn from the posterior distributions learned using the three different summary statistics learned via SASL, TLSL and DSSL. To do so, we first draw 500 parameter samples from the corresponding inferred approximate posterior distribution and simulate 500 data sets, each using a different parameter sample. This simulated dataset is considered as the predicted dataset from our inferred posterior distributions. In Fig 7, we plot the experimental data (solid line), 95% predictive credibility interval (shaded area) and the median prediction (dashed line) for SASL (red), DSSL (blue) and TLSL (green). The experimental data falls inside the 95% predictive credibility interval (PCI) for all the three summary learning approach, where DSSL producing the tightest PCI. For $\mathbb{N}_{agg-clust}$ and $\mathcal{S}_{agg-clust}$ the median prediction is closer to the true experimental data, whereas SASL and TLSL performs better for $\mathbb{N}_{platelet}$.

Further, to measure the deviation of the predicted dataset from the experimental data used for inference we use a Monte Carlo estimate of the energy score, which is a strictly proper scoring rule used to measure predictive performance of probabilistic predictions [45],

$$\text{Energy score} = 2\sum_i \|\boldsymbol{x}_i - \boldsymbol{x^0}\|_2^\beta - \sum_{i,j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^\beta, \tag{12}$$

where $\boldsymbol{x}_i$ is the i-th data simulated using a posterior sample, $\boldsymbol{x^0}$ is the experimental data used for inference, $\beta \in (0, 2)$. In Table 4, we report the energy score (fixing $\beta = 1$) computed correspondingly for the three inferential schemes using summary statistics learned via SASL, TLSL and DSSL for the three observed quantities $\mathbb{N}_{agg-clust}$, $\mathcal{S}_{agg-clust}$ and $\mathbb{N}_{platelet}$. The values in Table 4 are in agreement with Fig 7, allowing us to conclude that the proposed DSSL approach works better in predicting $\mathbb{N}_{agg-clust}$ and $\mathcal{S}_{agg-clust}$, whereas other two approaches perform better for the prediction of $\mathbb{N}_{platelet}$. This illustrates that the discriminative summary statistics learned by DSSL did not compromise in the overall predictive performance of the model, whereas it also encapsulates discriminative information between the three groups of patients/volunteers.

Table 4. Comparison of predictive performance of approximate ABC posterior learned via summary statistics learned using DSSL, SASL and TLSL approach based on energy score. Better predictive performance is measured by the smaller energy score values.

	DSSL	SASL	TLSL
$\mathbb{N}_{agg-clust}$	**2.71e+02**	1.08e+03	1.17e+03
$\mathcal{S}_{agg-clust}$	**3.47e+02**	2.16e+03	1.99e+03
$\mathbb{N}_{platelet}$	7.90e+04	4.16e+04	**2.46e+04**

https://doi.org/10.1371/journal.pcbi.1009910.t004

Maximum a posteriori estimate (MAP)

Given an observed dataset x^0, we want to estimate the corresponding θ. SABC inference scheme provides us with Z samples $(\theta_i)_{i=1}^Z$ from the ABC approximated posterior distribution $p(\theta|x^0)$ given the data for each patient. Given these samples we construct a smooth approximation of the posterior distribution (given data for a specific patient) of the parameters using Gaussian kernel density estimator with a bandwith equal to 0.45 and compute the mode of the smoothed posterior distribution using Nelder-Mead algorithm [46] as the estimate of parameters for each specific patient. This estimate will be considered as the MAP of the parameters. The Gaussian kernel density and Nelder-Mead algorithm were used as implemented in Python package 'scipy' [47].

In Fig 8 we illustrate the discriminative projection of the parameter space θ learnt using LMNN, by considering the MAP estimates, its second, third and fourth order polynomial expansion and the cross products between them. The tuning parameters for the LMNN algorithm were tuned as before to maximize the rand index [44] between the true patient clusters in the dataset and the clustering achieved using agglomerative hierarchical clustering with the Euclidean distance on the learned discriminative projection of parameter values from the MAP estimates. Euclidean distance between this learned parameter projection from the MAP estimates for different patient types were able to cluster the patients with 100% accuracy. This illustrates that we didn't lose any discriminative information inherent in the patient (volunteer) dataset by learning the MAP estimates, justifying our main claim that the values of the model parameters are precisely the information needed to assess various possible pathological conditions.

Here we further note that we could have used the discriminative summary statistics space or the discriminative projection of the parameter space correspondingly in Figs 6 and 8, which are not biologically meaningful, to construct the test to determine pathology in the patients with 100% accuracy. Instead we choose here to construct the test for pathology based on the biologically meaningful parameters of our stochastic deposition model, doing so our methodology somewhat looses accuracy but gains significant interpretability.

Conclusion

A numerical model of platelets deposition introduced in [25] was illustrated to successfully predict the platelets deposition and aggregation patterns in Impact-R device. [8] proposed an inferential scheme based on ABC using a distance based on expert knowledge to quantify the uncertainty of the model parameters of the model presented in [25] given data from a healthy patient. Due to the modeling of platelets reaching the bottom layers by a 1D diffusion, it was not possible to accurately calibrate the diffusion coefficient and the thickness of the boundary layers using the inferential scheme described in [8] and hence these quantities were manually tuned. To solve this problem, here we introduced a fully stochastic model which replaces the
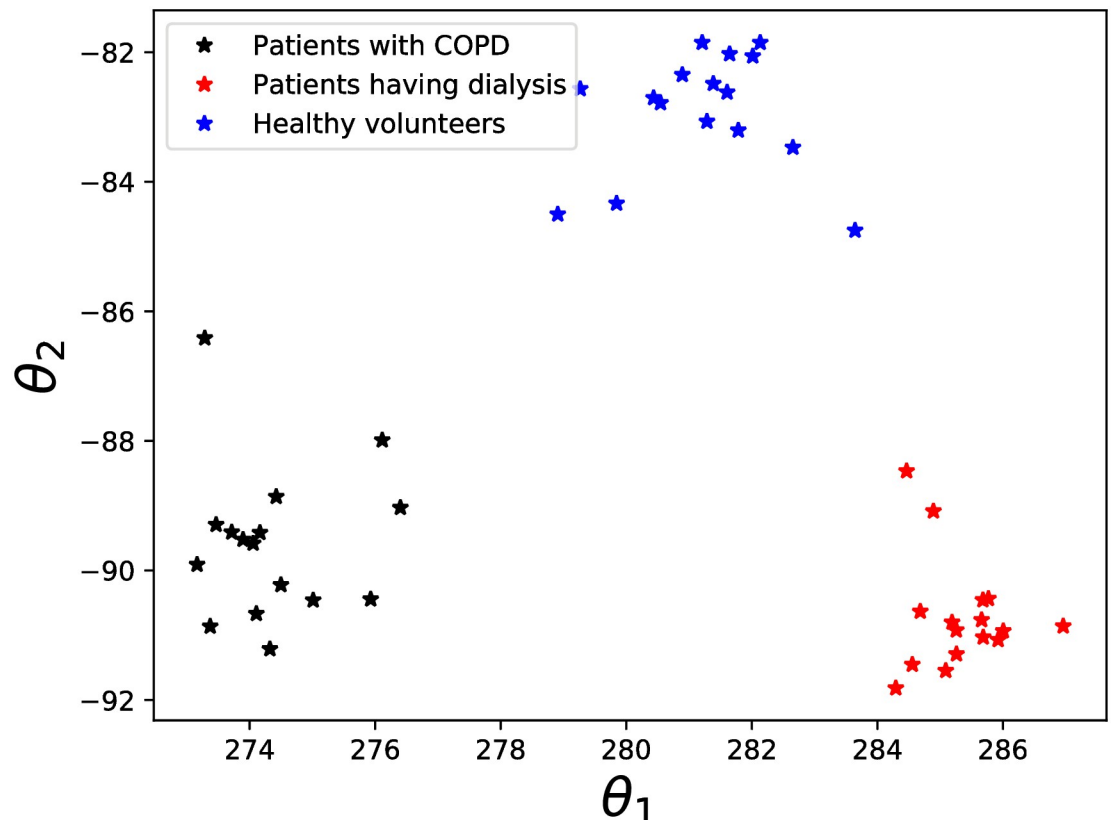
Fig 8. Discriminative projection of parameter space. The discriminative projection of the parameter space learned from the MAP estimates, in which the patients with COPD, patients having dialysis and healthy volunteers were accurately clustered illustrating that we do not lose any discriminative information inherent in the patient dataset while learning the MAP estimates.

https://doi.org/10.1371/journal.pcbi.1009910.g008

above parameters by the characteristic velocities of activated and non-activated platelets. We were able to estimate these velocities in addition to the other model parameters using the inferential scheme we propose in the present manuscript. Hence the proposed methodology completely removes the need to manually tune any parameters of the numerical model of platelets deposition.

Secondly, we adapt our inferential scheme to estimate parameters of the model which not only achieves good prediction performance but also can accommodate the discriminative information available in the dataset of different types of patients and volunteers. This is done by learning a summary statistic which is maximally discriminative between the three groups (healthy volunteers, patients with COPD and patients needing dialysis) considered and finally defining an Euclidean distance on this summary statistics space to use as distance between observed and simulated data in approximate Bayesian computation. We show that the three groups cluster accurately in this learned summary statistics space. This discriminative summary statistics was also able to get comparable predictive performance to the existing summary learning approaches when used in ABC to learn approximate posterior distribution of the model parameters given the data.

Finally, we evaluate the maximum a posteriori of the model parameters for each of the 48 patients by computing the mode of the joint approximate posterior distribution inferred by ABC. We notice estimated values of some of the parameters were able to distinguish between

different types of patients and healthy volunteers. We would like to note here that the original learned discriminative summary statistics could be used to differentiate between different types of patients and volunteers, but those summary statistics are not interpretable in a biological sense. Hence, our main contribution lies in this ability to estimate biologically meaningful parameters which can also discriminate between different types of patients. This may serve as an illustration to make some of the machine learning models used in biology interpretable.

Supporting information

S1 Appendix. Additional mathematical details. Details on Semi Automatic Summary Statistics Learning and Summary Statistics Learning by minimizing triplet loss.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Ritabrata Dutta, Karim Zouaoui Boudjeltia, Bastien Chopard.

Data curation: Karim Zouaoui Boudjeltia, Alexandre Rousseau, Daniel Ribeiro de Sousa, Jean-Marc Desmet, Alain Van Meerhaeghe, Bastien Chopard.

Formal analysis: Ritabrata Dutta.

Funding acquisition: Ritabrata Dutta, Karim Zouaoui Boudjeltia, Bastien Chopard.

Investigation: Ritabrata Dutta, Christos Kotsalos, Bastien Chopard.

Methodology: Ritabrata Dutta, Christos Kotsalos, Antonietta Mira, Bastien Chopard.

Project administration: Karim Zouaoui Boudjeltia, Bastien Chopard.

Resources: Antonietta Mira, Bastien Chopard.

Software: Ritabrata Dutta, Christos Kotsalos, Antonietta Mira, Bastien Chopard.

Validation: Ritabrata Dutta, Bastien Chopard.

Visualization: Ritabrata Dutta, Bastien Chopard.

Writing – original draft: Ritabrata Dutta.

Writing – review & editing: Ritabrata Dutta, Karim Zouaoui Boudjeltia, Christos Kotsalos, Antonietta Mira, Bastien Chopard.

References

1. Organization WH; 2015. http://www.who.int/mediacentre/factsheets/fs317/en/.

2. Breet NJ, van Werkum JW, Bouman HJ, Kelder JC, Ruven HJ, Bal ET, et al. Comparison of platelet function tests in predicting clinical outcome in patients undergoing coronary stent implantation. Jama. 2010; 303(8):754–762. https://doi.org/10.1001/jama.2010.181 PMID: 20179285

3. Picker SM. In-vitro assessment of platelet function. Transfus Apher Sci. 2011; 44:305–19. https://doi.org/10.1016/j.transci.2011.03.006 PMID: 21601530

4. Koltai K, Kesmarky G, Feher G, Tibold A, Toth K. Platelet aggregometry testing: Molecular mechanisms, techniques and clinical implications. International journal of molecular sciences. 2017; 18 (8):1803. https://doi.org/10.3390/ijms18081803 PMID: 28820484

5. Chopard B, de Sousa DR, Lätt J, Mountrakis L, Dubois F, Yourassowsky C, et al. A physical description of the adhesion and aggregation of platelets. Royal Society open science. 2017; 4(4):170219. https://doi.org/10.1098/rsos.170219 PMID: 28484643

6. Shenkman B, Einav Y, Salomon O, Varon D, Savion N. Testing agonist-induced platelet aggregation by the Impact-R [Cone and plate (let) analyzer (CPA)]. Platelets. 2008; 19(6):440–446. https://doi.org/10.1080/09537100802082256 PMID: 18925512

7. Lintusaari J, Gutmann MU, Dutta R, Kaski S, Corander J. Fundamentals and recent developments in approximate Bayesian computation. Systematic biology. 2017; 66(1):e66–e82. https://doi.org/10.1093/sysbio/syw077 PMID: 28175922

8. Dutta R, Chopard B, Lätt J, Dubois F, Zouaoui Boudjeltia K, Mira A. Parameter estimation of platelets deposition: Approximate Bayesian computation with high performance computing. Frontiers in physiology. 2018; 9:1128. https://doi.org/10.3389/fphys.2018.01128 PMID: 30177886

9. Suárez JL, García S, Herrera F. A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms and Software. arXiv preprint arXiv:181205944. 2018.

10. Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research. 2009; 10(2).

11. Pacchiardi L, Künzli P, Schoengens M, Chopard B, Dutta R. Distance-learning For Approximate Bayesian Computation To Model a Volcanic Eruption. Sankhya B. 2020; p. 1–30.

12. Berger JO. Statistical decision theory and Bayesian analysis. Springer Science & Business Media; 2013.

13. Van der Vaart AW. Asymptotic statistics. vol. 3. Cambridge university press; 2000.

14. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995; 57(1):289–300.

15. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association. 1952; 47(260):583–621. https://doi.org/10.1080/01621459.1952.10483441

16. Mourikis P, Helten C, Dannenberg L, Hohlfeld T, Stegbauer J, Petzold T, et al. Platelet reactivity in patients with chronic kidney disease and hemodialysis. Journal of Thrombosis and Thrombolysis. 2020; 49(1):168–172. https://doi.org/10.1007/s11239-019-01983-x PMID: 31691067

17. Thijs A, Nanayakkara P, Ter Wee P, Huijgens P, Van Guldener C, Stehouwer C. Mild-to-moderate renal impairment is associated with platelet activation: a cross-sectional study. Clinical nephrology. 2008; 70(4):325. PMID: 18826858

18. Mallah H, Ball S, Sekhon J, Parmar K, Nugent K. Platelets in chronic obstructive pulmonary disease: An update on pathophysiology and implications for antiplatelet therapy. Respiratory Medicine. 2020; p. 106098. https://doi.org/10.1016/j.rmed.2020.106098 PMID: 32777683

19. Malerba M, Olivini A, Radaeli A, Ricciardolo FLM, Clini E. Platelet activation and cardiovascular comorbidities in patients with chronic obstructive pulmonary disease. Current medical research and opinion. 2016; 32(5):885–891. https://doi.org/10.1185/03007995.2016.1149054 PMID: 26824631

20. Bujak K, Wasilewski J, Osadnik T, Jonczyk S, Kołodziejska A, Gierlotka M, et al. The prognostic role of red blood cell distribution width in coronary artery disease: a review of the pathophysiology. Disease markers. 2015; 2015. https://doi.org/10.1155/2015/824624 PMID: 26379362

21. Piagnerelli M, Boudjeltia KZ, Brohée D, Vereerstraeten A, Piro P, Vincent JL, et al. Assessment of erythrocyte shape by flow cytometry techniques. Journal of clinical pathology. 2007; 60(5):549–554. https://doi.org/10.1136/jcp.2006.037523 PMID: 16775118

22. Zouaoui Boudjeltia K, Kotsalos C, de Sousa DR, Rousseau A, Lelubre C, Sartenaer O, et al. Spherization of red blood cells and platelet margination in COPD patients. Annals of the New York Academy of Sciences. 2020;. PMID: 33009705

23. Shreffler J, Huecker MR. Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios. 2020.

24. Chopard B, Ribeiro de Sousa D, Latt J, Dubois F, Yourassowsky C, Van Antwerpen P, et al. A physical description of the adhesion and aggregation of platelets. ArXiv e-prints. 2015.

25. Chopard B, de Sousa DR, Lätt J, Mountrakis L, Dubois F, Yourassowsky C, et al. A physical description of the adhesion and aggregation of platelets. Royal Society Open Science. 2017; 4(4):170219. https://doi.org/10.1098/rsos.170219 PMID: 28484643

26. Bian X, Kim C, Karniadakis GE. 111 years of Brownian motion. Soft Matter. 2016; 12:6331–6346. https://doi.org/10.1039/c6sm01153e PMID: 27396746

27. Kotsalos C, Latt J, Beny J, Chopard B. Digital Blood in Massively Parallel CPU/GPU Systems for the Study of Platelet Transport. Interface Focus (Royal Society). 2020;(11):20190116.

28. Beaumont MA. Approximate Bayesian computation in evolution and ecology. Annual review of ecology, evolution, and systematics. 2010; 41:379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621

29. Drovandi CC, Pettitt AN. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. Biometrics. 2011; 67(1):225–233. https://doi.org/10.1111/j.1541-0420.2010.01410.x PMID: 20345496

30. Del Moral P, Doucet A, Jasra A. An adaptive sequential Monte Carlo method for approximate Bayesian computation. Statistics and Computing. 2012; 22(5):1009–1020. https://doi.org/10.1007/s11222-011-9271-y

31. Lenormand M, Jabot F, Deffuant G. Adaptive approximate Bayesian computation for complex models. Computational Statistics. 2013; 28(6):2777–2796. https://doi.org/10.1007/s00180-013-0428-3

32. Albert C, Hans RK, Scheidegger A. A simulated annealing approach to approximate Bayesian computations. Statistics and Computing. 2015; 25:1217–1232. https://doi.org/10.1007/s11222-014-9507-8

33. Dutta R, Schoengens M, Ummadisingu A, Onnela JP, Mira A. ABCpy: A High-Performance Computing Perspective to Approximate Bayesian Computation. Journal of Statistical Software. 2021; 100(7):1–38. https://doi.org/10.18637/jss.v100.i07

34. Dutta R, Schoengens M, Onnela JP, Mira A. ABCpy: A user-friendly, extensible, and parallel library for Approximate Bayesian Computation. In: Proceedings of the Platform for Advanced Scientific Computing Conference. ACM; 2017.

35. Gutmann MU, Dutta R, Kaski S, Corander J. Likelihood-free inference via classification. Statistics and Computing. 2017; p. 1–15. https://doi.org/10.1007/s11222-017-9738-6 PMID: 31997856

36. Jiang B. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In: International conference on artificial intelligence and statistics. PMLR; 2018. p. 1711–1721.

37. Park M, Jitkrittum W, Sejdinovic D. K2-ABC: Approximate Bayesian computation with kernel embeddings. In: Artificial intelligence and statistics. PMLR; 2016. p. 398–407.

38. Bernton E, Jacob PE, Gerber M, Robert CP. Approximate Bayesian computation with the Wasserstein distance. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2019; 81(2):235–269. https://doi.org/10.1111/rssb.12312

39. Didelot X, Everitt RG, Johansen AM, Lawson DJ, others. Likelihood-free estimation of model evidence. Bayesian Analysis. 2011; 6(1):49–76. https://doi.org/10.1214/11-BA602

40. Csilléry K, Blum MG, Gaggiotti OE, François O. Approximate Bayesian computation (ABC) in practice. Trends in Ecology & Evolution. 2010; 25(7):410–418. https://doi.org/10.1016/j.tree.2010.04.001 PMID: 20488578

41. Fearnhead P, Prangle D. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation [with Discussion]. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2012; 74(3):419–474. https://doi.org/10.1111/j.1467-9868.2011.01010.x

42. Jiang B, Wu Ty, Zheng C, Wong WH. Learning summary statistic for approximate Bayesian computation via deep neural network. Statistica Sinica. 2017; p. 1595–1618.

43. De Vazelhes W, Carey C, Tang Y, Vauquier N, Bellet A. metric-learn: Metric learning algorithms in python. Journal of Machine Learning Research. 2020; 21(138):1–6.

44. Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association. 1971; 66(336):846–850. https://doi.org/10.1080/01621459.1971.10482356

45. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association. 2007; 102(477):359–378. https://doi.org/10.1198/016214506000001437

46. Gao F, Han L. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. Computational Optimization and Applications. 2012; 51(1):259–277. https://doi.org/10.1007/s10589-010-9329-3

47. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020; 17:261–272. https://doi.org/10.1038/s41592-019-0686-2 PMID: 32015543