Statistical models for fever forecasting based on advanced body temperature monitoring

Jorge Jordan, Pau Miro–Martinez, Borja Vargas, Manuel Varela–Entrecanales, David Cuesta–Frau

PII: S0883-9441(16)30532-9
DOI: doi: 10.1016/j.jcrc.2016.09.013
Reference: YJCRC 52283

To appear in: *Journal of Critical Care*

Please cite this article as: Jordan Jorge, Miro–Martinez Pau, Vargas Borja, Varela–Entrecanales Manuel, Cuesta–Frau David, Statistical models for fever forecasting based on advanced body temperature monitoring, *Journal of Critical Care* (2016), doi: 10.1016/j.jcrc.2016.09.013

Statistical models for fever forecasting based on advanced body temperature monitoring

Jorge Jordan and Pau Miro–Martinez
Department of Statistics, Polytechnic University of Valencia, Alcoi Campus, Plaza Ferrandiz y Carbonell, 2, 03801 Alcoi (Spain) E-mail: jorjornu@eio.upv.es

Borja Vargas
Biomedical Science Faculty, European University of Madrid, Villaviciosa de Odon, 28670 Madrid (Spain)

Manuel Varela–Entrecanales
Department of Internal Medicine, Teaching Hospital of Mostoles, Madrid (Spain)

David Cuesta–Frau
Technological Institute of Informatics, Polytechnic University of Valencia, Alcoi Campus,
Plaza Ferrandiz y Carbonell, 2, 03801 Alcoi (Spain)

Keywords Thermometry · Fever · Temperature monitoring · Infectious diseases · Entropy · ApEn

Abstract Body temperature monitoring provides healthcarers with key clinical information about the physiological status of patients. Temperature readings are taken periodically to detect febrile episodes and consequently implement the appropriate medical countermeasures. However, fever is often difficult to assess at early stages, or remains undetected until the next reading, probably a few hours later. The objective of this paper is to develop a statistical model to forecast fever before a temperature threshold is exceeded to improve the therapeutic approach to the subjects involved. To this end, temperature series of nine patients admitted to a general Internal Medicine ward were obtained with a continuous monitoring holter device, collecting measurements of peripheral and core temperature once per minute. These series were used to develop different statistical models that could quantify the probability of having a fever spike in the following 60 minutes. A validation series was collected to assess the accuracy of the models. Finally, the results were compared with the analysis of some series by experienced clinicians. Two different models were developed: a logistic regression model and a linear discrimination analysis model. Both of them exhibited a fever peak forecasting accuracy above 84%. When compared with experts assessment, both models identified 35 out of 36 fever spikes (97.2%). The models proposed are highly accurate in forecasting the appearance of fever spikes within a short period of time in patients with suspected or confirmed febrile related illnesses.

1 Introduction

Body temperature is universally considered a crucial vital constant, and is systematically recorded and analyzed in every admitted patient. Its clinical usefulness is unquestionable. However, clinical thermometry is heavily hampered by several conceptual problems: 1) body temperature is considered a “constant”, and as such, measured at low frequency (e.g. three times a day, or on every shift). This concept is arguably an offspring from the classical humour theory (and temperature would be the equivalent to the “heat” element). 2) We tend to assume that the body has a single temperature, which is kept constant through a cybernetic mechanism regulated by the hypothalamic thermostat. 3) Fever parallels the course of (mainly) infectious diseases. 4) There is a “normal limit” of temperature, above which the patient should be considered febrile.

Each of these assumptions is hardly tenable. Instead, body temperature is an ever–moving, fine–tuned equilibrium between heat production and heat dissipation. The concept of a central hypothalamic centre is being substituted by a network of afferent and efferent overlapping loops [1]. One of the central thermoregulating mechanisms is a tight control of the amount of blood being circulated through cutaneous capillaries vs. the amount of blood being short-circuited through deeper arterio–venous shunts. The balance between these two arms is under control of the autonomic system, and regulates the amount of heat being transferred to the environment. Thus, the gradient between central and peripheral temperature displays the “heat–conserving” or “heat–dissipating” mode of the body at each moment.

Furthermore, building a fever is not an immediate process. It is a painstaking, metabolically demanding process that requires several steps from the initial production of lipopolysaccharides and other pyrogenic substances by infectious organisms, through macrophages activation, interleukin liberation and activation of the afferent and efferent thermoregulating loops before body temperature rises. This may take time, as proved by the fact that bacterial blood counts are high in the pre–febrile state [2,3].

Finally, while clinical decisions must be made and thus we need some kind of threshold, it is probably naive to assume there is a pre-fixed “red line” separating febrile from afebrile patients. Where to put that clinical threshold varies depending on the patient and the context, and is arguably a typical example of clinical judgement. Obviously, admitting the blurred limits of temperature does not question the existence of fever. Fever is an indisputable clinical phenomenon that, while originally a defence mechanism, may be a trial for frail or unstable patients.

Being able to predict the development of fever in a specific time–lapse may have important consequences. It may allow obtaining blood cultures when the bacterial count is peaking, thus increasing the diagnostic yield. It may also prompt preventive or therapeutic measures to avoid or curtail the febrile episode in patients in which a fever may be especially undesirable [4]. In this context, the term fever may have two different meanings:

- A core temperature above a predefined threshold (e.g. 38°C).

- A temperature profile that a set of qualified clinicians would consider clinically relevant, and/or at which they would decide to obtain blood cultures.

Our group has already described the usefulness of a holter device that allows continuous monitoring of central and peripheral temperatures to identify fever spikes. Some of them were overlooked by conventional measurements [5], therefore suggesting one advantage of performing a continuous surveillance of temperature. In addition, this device can measure both central and surface temperatures at the same time, which in turn allows to approximate to cutaneous blood flow regulation, a key factor in the appearance of fever.

We have already demonstrated that complexity analysis of temperature curves reflects the severity of patients admitted to the critical care unit and it is a prognostic marker [6]. In this setting, complexity analysis may provide important information about the processes governing the regulation of body temperature during fever.

Based on these previous studies, we hypothesized that it could be possible to forecast the appearance of fever spikes in patients diagnosed or suspected to suffer an infection with the use of a holter device that monitors both central and skin surface temperatures. Therefore, the present study had three successive objectives:

1. To develop a model that, through the real–time recording and analysis ofthe co–evolution of central and peripheral temperature of a patient, would be able to foresee and alert clinicians on the development of fever in the following 60 minutes.
2. To validate the results of such model on a different sample of patients.
3. To compare the results of such model with the clinical judgement of a set of qualified physicians.

2 Methods

Two modelling schemes were addressed: linear discrimination analysis (LDA) and logistic regression (LR). The variables were characterized with a univariate or a bivariate analysis in order to choose the most appropriate construction techniques for the models. All explanatory variables chosen were quantitative, whereas the independent variable was qualitative (Fever peak anticipated or not). The variables can have different means and medians under these schemes.
Although some variables may not follow the normal distribution, LDA can be applied anyway if model adjustment is proven to be correct [12]. LR is not affected by this requirement.

The presence of multicollinearity among the explanatory variables has to be evaluated in order to remove those variables that could cause redundancy. Three tests were performed consecutively to check for multicollinearity in the method proposed:

1. The correlation matrix, where a high correlation is detected if the coefficients R_{ij} satisfy $|R_{ij}| \geq 0.7$.
2. The inverse correlation matrix, where a high correlation was said to bepresent when $R_{ii}^{-1} > 10.$
3. The method of Belsley, Kuh and Welsch [10]. In this test, index conditioning (IC) is computed from the eigenvalues λ of the correlation matrix as $\mathrm{IC} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$, where IC values greater than 10 indicate multicollinearity problems.

In our case, the results of multicollinearity analyses showed that there was a significant relation among the explanatory variables. To solve this problem, we chose a method of gradual variable elimination [11]. The models were optimised by selecting the variables in their significance order. The resulting models were finally validated to assess their predictive accuracy as described in Section 4.

2.1 Model variables

A fever peak is considered to occur when central temperature is higher than 38°C, with at least one measurement lower than 37.5°C in the previous 30 min. In accordance with this definition, a dichotomic output variable termed "Signal" was defined, with points in the 60 minutes prior to this peak assigned a value of 1, and the rest of the points being 0. An example of the behaviour of the Signal variable is depicted in Fig. 1. During the next 120 minutes after a peak, no further peaks can be forecast to avoid redundant information about the subject status. Namely, the models are not retriggerable.

The input variables for the development of the predictive models were:

- Central temperature (T_c): measurement of temperature in the external auditory canal (EAC). It was assumed that this temperature is a surrogate of the body's internal temperature (core), although some differences may exist. Theoretically, T_c is a key factor for fever forecasting, as a temperature held above 37°C for some minutes is very likely to be related with a fever spike.
- Peripheral temperature (T_p): measurement of skin temperature on the anterior surface of the forearm. Skin temperature is highly dependent on environmental temperature, and it is not the same throughout the body surface, but it is useful for measuring the gradient with T_c and make an approach to heat loss, a key factor in thermoregulation and fever.

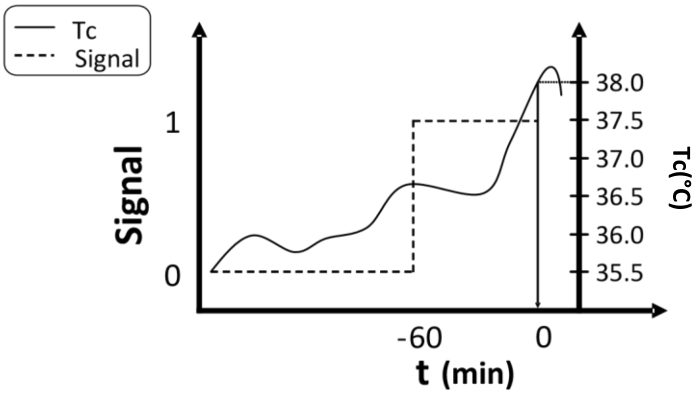


Fig. 1 Signal variable behaviour. Points in the 60 minutes before a fever spike take the value of 1, otherwise the value assigned is 0.

- Gradient between T_c and T_p (G_{cp}): the difference between each measurement of T_c and T_p. From a pathophysiological perspective, fever is preceded by a rise in this gradient, as an indicator of cutaneous vasoconstriction to keep the body warm.
- Gradient mean ($\bar{G}_{cp}$): the average value of G_{cp} during the last 60 minutes.
- Correlation coefficient between Tc and Tp ($\bar{C}_{cp}$): the mean of the correlation coefficient between T_c and T_p in the last 60 minutes.
- Correlation coefficient between Tc and Grad ($\bar{C}_{cGcp}$): the mean of the correlation coefficient between T_c and the gradient in the last 60 minutes.
- Correlation gradient (G_C): the difference between $\bar{C}_{cp}$ at time n and $\bar{C}_{cp}$ at time $n - 30$. This value was estimated because a sharp fall in $\bar{C}_{cp}$ was often found in the minutes preceding a fever peak.
- Approximate Entropy (ApEn) of T_c, T_p and G_{cp} (A_{Tc}, A_{Tp}, A_{Gcp}). ApEn evaluates the predictability of a time series, measuring to what extent a certain pattern predicts the ensuing points. Measures of ApEn of T_c, T_p and G_{cp} were considered to provide information about changes in sharp regulation of body temperature at the first stages of fever. Details about ApEn and related statistics may be found elsewhere [8]. In this work, ApEn was calculated with N = 120, m = 1 and r = 0.2∗SD.
- Cross–ApEn of T_c and T_p (CA): Cross–ApEn is a parameter related with ApEn that evaluates the synchrony of two different time series. Here, N = 120, m = 1 and r = 0.2∗SD.

3 Experiments

3.1 Training experimental set

These temperature time series were collected among patients admitted to the Internal Medicine ward of a teaching hospital in Mostoles, Madrid, from 2008 to 2010. The sample included 62 patients who had had a temperature measurement above 38°C the day before they were monitored. They were also required to be over 18 and under 85 years of age and to have been admitted for less than a week. Central and peripheral temperatures were recorded with two sensors placed in the external auditory canal (Mon–a–Therm Tympanic Temperature Probe, Covidien) for central temperature and in the cubital aspect of the forearm (Mon–a–Therm Skin Temperature Probe, Covidien) for peripheral temperature, and stored every minute for 24h using a temperature Holter device (TherCom, Innovatec) described elsewhere [7]. The results of different analyses performed on those series have been published in [5]. Written informed consent was obtained from each participant before recruitment and monitorization.

Since analyses of entropy derived techniques require perfectly accomplished time series, temperature recordings from patients in the sample described above were reviewed, and eventually 9 series (for a total of 8325 temperature readings) were chosen for the development of predictive models: 6 men and 3 women, with a median age of 56 years (range 32 to 81). They were considered specifically suitable for this purpose since they had no signal losses or disconnections.

3.2 Validation experimental set

The validation series were collected in the same manner as for the training set. A total of 14 patients were recruited between May and July 2013, following the same criteria, and they were monitored with the same devices and sensors. Temperature recordings were reviewed and 8 of them were considered suitable to be analyzed with the predictive models for validation, for a total of 7486 temperature readings. This set included 6 men and 2 women, with a median age of 59 years (range 31 to 63). Written informed consent was obtained from each subject.

Accuracy of each model was evaluated by comparison of the predicted value of the variable "Signal" with its real value for each point in the whole sample. A contingency table was built with these results, and global accuracy was calculated as the total percentage of correct forecasts. Models with the highest global accuracy were chosen.

3.3 Experts validation

Five specialists in Internal Medicine of our hospital were chosen to evaluate some of the temperature series used for the development of the predictive models. Three of them work in a general Internal Medicine ward and two of them in an Infectious Diseases ward. All of them had more than ten years of clinical experience.

The survey included 30 temperature series, with 6 of them repeated twice. The physicians were requested to mark the points where they considered a fever

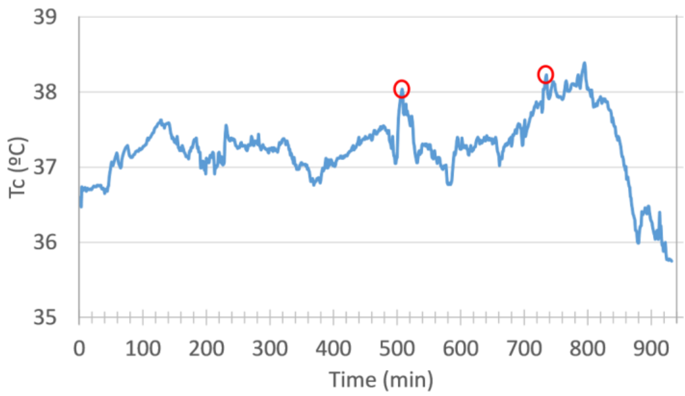


Fig. 2 Temperature series example for central temperature (T_c). Circles represent the fever peaks detected by the medical specialists.

peak to begin on the graphics. It was considered that an agreement existed when three or more physicians marked the same point. Fig. 2 depicts an example of a body temperature record, with circles indicating the beginning of fever spikes as considered by the experts.

These experts identified 42 fever peaks in the 30 series that they analyzed, based on the criteria previously established (agreement between three or more of them). A total of 6 of these peaks could not be used for comparison with the predictive models because they occurred less than 120 minutes after the previous one, and those time windows had been excluded from the predictive models, as previously explained.

4 Results

Threshold values were set at 0.880 for the LDA model and 0.160 for the LR model, since they yielded the highest accuracy. Fig. 3 show the receiver operator characteristic (ROC) for the LDA and RL with the area under curve (AUC). The AUC for both models are very near to 1, and are considered very good models to predict. Values above the threshold were considered as predictive of a fever spike occurring in the next 60 minutes. With the cutoff values previously defined, the LDA model correctly classified 84.76% of the training set, and the LR model, a 84.58%. Among points with a value of "0" for Signal, classification was correct in 84.61% of the cases for the LR (6165 out of 7286) and in 84.77% of the cases for the LDA (6176 out of 7286). Among points with a value of "1", the LR model correctly classified 876 points out of 1039 (84.58%) and the LDA model 880 out of 1039 (84.7%) (see Table 1). Mean anticipation time to a fever peak was 82 minutes for the LDA model (SD 44 minutes) and 84 minutes for the LR model (SD 44 minutes).

As for the validation set, the RL model correctly classified 6695 of 7486 points (global accuracy of 89.43%). The LDA correctly classified 6684 mea-

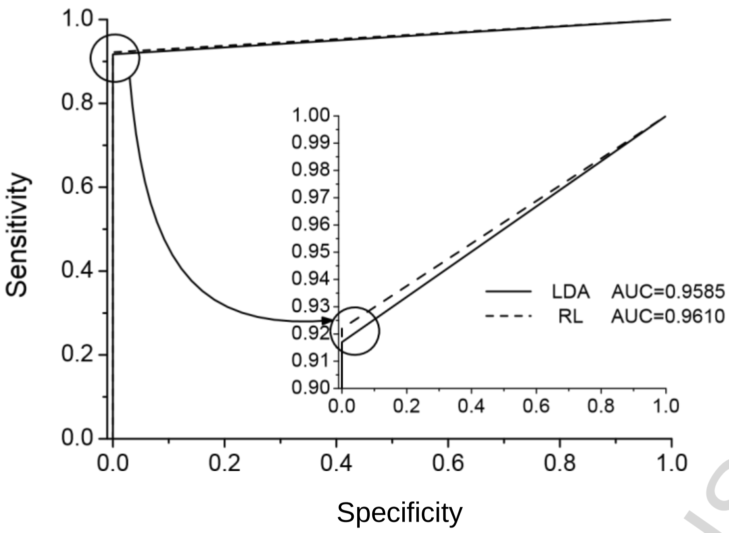


Fig. 3 ROC curve for LDA and RL with the area under curve (AUC). The LDA curve appears in solid line, and the RL in dash line.

Table 1 Accuracy of LDA and LR models over the training experimental set.

			Forecast					
			LDA			RL		
			0	1	Total	0	1	Total
	Count	0	6176	1110	7286	6165	1121	7286
		1	159	880	1039	163	876	1039
Observed		0	84.77	15.23	100	84.61	15.39	100
	%	1	15.30	84.70	100	15.69	84.58	100
	Global accuracy				84.76%			84.58%

surements (global accuracy of 89.29%). Among points with a value of "0" for Signal, classification was correct in 92.06% of the cases for the LR (6457 out of 7014) and in 91.93% of the cases for the LDA (6448 out of 7014). Among points with a value of "1", the LR model correctly classified 238 points out of 472 (50.42%) and the LDA model 236 out of 472 (50%). Both the LDA and the LR models identified 35 out of the remaining 36 peaks, for a 97.2% sensitivity. In the two series where no fever spikes were identified by doctors, the models performed properly, and no false positives took place. Mean anticipation time to a fever peak was 49 minutes for the LDA model (SD 30.17 minutes) and 51.42 minutes for the LR model (SD 30.35 minutes). An example of the performance of the LDA predictive model on one series is shown in Fig. 4.

The hypotheses of each model were checked using the statistical software SPSS 22 of IBM Enterprise. The LDA model performed correctly, according to the Wilks' Lambda test (p = 0.000), and all the explanatory variables were useful to discriminate between the minutes before a fever spike and the rest of the series, according to the Wilks' Lambda test (p < 0.05 in all cases). The normality of explanatory variables was assumed because the number of obser-

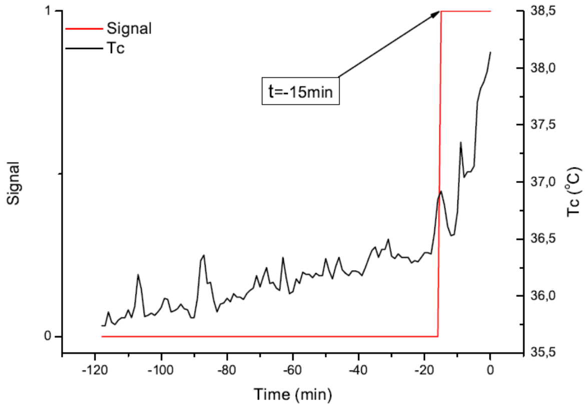


Fig. 4 Performance of the LDA predictive model. The red line represents the value of the variable "Signal", calculated by this model. In this case, the model forecast a fever spike 15 minutes before it appeared, with a core body temperature measurement of 36.9°C.

vations was very high (greater than 8000). Covariance arrays were different according to test box (p = 0.000), which may adversely affect the estimation of parameters as previously stated [9]. The LR model also performed correctly according to the Hosmer and Lemeshow test (p–value = 0.699), and the explanatory variables were also useful for the model according to the Wald test (p–value < 0.05 for each variable).

Several LDA and LR models were developed and compared. The most accurate versions were:

$$D = -56.759 + 1.557 * T_c - 0.385 * G_{cp} - 0.834 * \bar{C}_{cp} - 0.344 * \bar{C}_{cGcp} + \\ +0.374 * G_C + 0.658 * A_{Tc} + 0.422 * A_{Gcp} \quad (1)$$

for the LDA case, and:

$$Ln\left(\frac{p_i}{1-p_i}\right) = f(x) \quad (2)$$

where

$$f(x) = -169.373 + 4.529 * Tc - 1.035 * G_{cp} - 2.058 * \bar{C}_{cp} - 0.721 * \bar{C}_{cGcp} + \\ +0.815 * G_C + 1.621 * A_{Tc} + 1.158 * A_{Tp} \quad (3)$$

for the LR model.

5 Discussion

Model development was based on each temperature recording, and predictive accuracy was estimated on the prediction made for the dependent variable

(Signal) at each point. Following this criteria, both models display very high global accuracy rates (above 84% for the original series and above 89% for the validation series). When these results are assessed in detail, the predictive accuracy was high for points with values of "0" (those that were more than 60 minutes before a fever peak) and "1" (those that are in the 60 minutes previous to a fever peak) for the dependent variable in the original series, whereas in the validation series accuracy rates for points with a value of "1" was around 50%, which could be considered as a very low sensitivity. This is due to the interpretation of the results taking into account each temperature measurement as an individual point. Conversely, if each fever peak is assessed as a whole, both models were able to forecast all fever peaks in the original series and the validation series, with those low sensitivity rates indicating that sometimes the fever was anticipated by 10 or 20 minutes. The same problem could be considered regarding incorrect prediction of "0" values (what could be interpreted as false positives). Since the prediction range was limited to 60 minutes before a T_c measurement above 38°C, values of "1" that were more than 60 minutes apart from a fever peak were considered as a false positive in the contingency table.

6 Conclusion

We described a method to forecast fever in this paper. It is based on statistical modelling using core and peripheral body temperature data, and related parameters. The main finding of our study is that both models introduced were highly accurate in forecasting the appearance of fever spikes, with accuracy rates above 84%.

Nevertheless, some caveats must be mentioned. First of all, we realize that the definitions we have applied are some way arbitrary. We think this is a common problem when measuring body temperature, and it has not been solved by now. The very definition of fever as a rise in central temperature above 38°C could be a misunderstanding. We accepted it as the threshold of fever, regardless of the time of the day, the use of antipyretic drugs or the clinical status, and it was compulsory to follow this criterion to standardize the results.

Secondly, the models require a highly reliable time–series, as the measurements of complexity parameters are highly dependent on the temporal evolution of the signal. Thus, any disconnection of the sensors could make a period of about two hours useless for further analysis. Unfortunately, we faced this problem too often during the sample collection, and some series were removed from the final analysis because they were considered unsuitable.

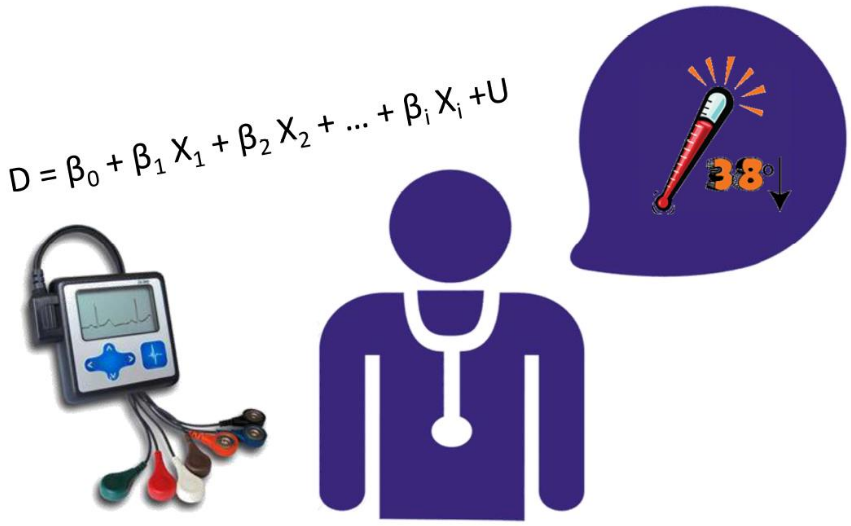
For these reasons, we decided to perform a process of validation by experts. As explained above, in this stage we compared the detection of fever peaks by experts with the predictions of the models, considering each fever peak as a whole. With this criteria, both models identified 35 out of 36 fever peaks defined by experts, and no false positive was observed. Although in our opinion these results are much more significant from a clinical perspective, we strongly believe that the most important criteria to evaluate the usefulness of the models would be their clinical application to particular fields, which should be elucidated in the future.

Acknowledgements No external financial support was obtained.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Romanovsky A., Thermoregulation: some concepts have changed. Functional architectureof the thermoregulatory system, Am. J. Physiol. Regul. Integr. Comp. Physiol., 292, 37–46 (2007)
2. Bennett I.L., Beeson R.B., Bacteremia: A consideration of some experimental and clinicalaspects, Yale Journal of Biology and Medicine, 262, 241–262 (1954)
3. Mylotte J.M., Tayara A., Blood cultures: Clinical aspects and controversies, Eur. J. Clin.Microbiol. Infect. Dis., 19, 157–63 (2000)
4. Dinarello C.A., Porat R. Chapter 16. Fever and Hyperthermia. In: Longo DL, Fauci AS,Kasper DL, Hauser SL, Jameson JL, Loscalzo J, editors. Harrison's Principles of Internal Medicine, 18e. McGraw-Hill, New York, NY (2012)
5. Varela–Entrecanales M., Ruiz–Esteban R., Martinez–Nicolas A., Cuervo–Arango J.A.,Barros C., Delgado E., Catching the spike and tracking the flow: Holter-temperature monitoring in patients admitted in a general internal medicine ward, Int. J. Clin. Pract., 65, 1283–1288 (2011)
6. Varela–Entrecanales M., Churruca J., Gonzalez A., Martin A., Ode J., Galdos P., Temperature curve complexity predicts survival in critically ill patients, Am. J. Respir. Crit. Care Med., 174(3), 290–298 (2006)
7. Cuesta-Frau D., Varela–Entrecanales M., Aboy M., Miro–Martinez P., Description of aportable wireless device for high–frequency body temperature acquisition and analysis, Sensors, 9, 7648–7663 (2009)
8. Pincus S.M., Approximate entropy as a measure of system complexity, Proc. Natl. Acad.Sci. USA, 88(6), 2297–301 (1991)
9. Tabachnick B. G., Fidell L. S., Using Multivariate Statistics (5th Edition), Allyn &Bacon, Inc., Needham Heights, MA, USA (2006)
10. Belsley D.A., Kuh E., Welsch R.E., Regression diagnostics: Identifying influential dataand sources of collinearity, Wiley–Interscience (2005)
11. Yoo W., Mayberry R., Bae S., Singh K., He Q., Lillard J.W., A study of effects of multicollinearity in the multivariable analysis. International journal of applied science and technology, 4(5):9–19 (2014)
12. Gessner G., Malhotra N.K., Kamakura W.A., Zmijewski M.E., Estimating models withbinary dependent variables: Some theoretical and empirical observations, Journal of business research, 16(1), 49–65 (1988)


$D = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i + U$
38°

Graphical Abstract