Clinically relevant features for predicting the severity of surgical site infections

Ahcène Boubekki[1], Jonas Nordhaug Myhre[1,4], Luigi Tommaso Luppino[1], Karl Øyvind Mikalsen[1,2,3], Arthur Revhaug[2,3], and Robert Jenssen[1]

***Abstract*—Surgical site infections are hospital-acquired infections resulting in severe risk for patients and significantly increased costs for healthcare providers. In this work, we show how to leverage irregularly sampled preoperative blood tests to predict, on the day of surgery, a future surgical site infection and its severity. Our dataset is extracted from the electronic health records of patients who underwent gastrointestinal surgery and developed either deep, shallow or no infection. We represent the patients using the concentrations of fourteen common blood components collected over the four weeks preceding the surgery partitioned into six time windows. A gradient boosting based classifier trained on our new set of features reports an AUROC of 0.991 for predicting a postoperative infection and and AUROC of 0.937 for classifying the severity of the infection. Further analyses support the clinical relevance of our approach as the most important features describe the nutritional status and the liver function over the two weeks prior to surgery.**

***Index Terms*—surgical site infection, machine learning,**

I. Introduction

SURGICAL Site Infections (SSI) are some of the most common hospital-acquired infections [1], representing up to 30% of the total amount of such infections [2], [3], [4]. At the same time, SSIs are considered some of the most preventable kinds of infections [5]. SSIs can be divided into different types depending on the anatomical location of the infection [6] and associated care: *Superficial* infections can be treated with surgical debridement and sometimes antibiotics, whereas *deep* infections are more complex and may require another surgery (e.g. laparatomies), another medical procedure (e.g. percutaneous drainage), or intravenous antibiotics. Distinguishing the severity of the infection is thus pertinent from a clinical, practical and economical point of view [7], [8], [9]. Established risk factors are age, overweight, smoking, type of surgery and cancer [10], [11]. Along with an increased mortality rate (3%), an SSI can also prolong the postoperative hospital stay by as much as two weeks [12] or require a readmission [13] which is estimated to cost up to 27,000 USD [7]. It is clear that reducing the risk of postoperative complications stemming from SSIs, especially from the deep infections, will be of great benefit for the patients, the healthcare systems and the society in general.

Blood tests present several advantages as they are already part of clinical routines and can be performed with low burden for the patient. A wide variety of statistical methods have been studied to leverage this type of data [14], [8], [15], [16], [9], [17]. Several recent studies have focused on the analysis and prediction of SSI from *blood test values* collected before and/or after the surgery [18], [19], [20], [21]. However, to our knowledge, no previous work has focus only on preoperative data and stratified the analysis based on infection depth, despite the difference being clinically relevant [8], [9]. This is certainly due to a lack of relevant public baseline data. We are basing our study on a data set from the University Hospital of North Norway released for the Knowledge Discovery and Data Mining competition organized by the American Medical Informatics Association in 2016. It was also used by Soguero-Ruiz *et al.* [8] and Kocbek *et al.* [21].

To solve our specific task, we propose a new feature extraction pipeline: 1) To have practical clinical impact, prediction models and support tools need to serve clinicians *on the day* of surgery. Therefore, we do not include post-surgery exams; 2) To allow for flexible pre-surgery monitoring, we rely on weekly averages over the three weeks preceding the one of the surgery. Averages make features less sensitive to missing data and temporal variations (samples taken at the beginning or at the end of the week); 3) To capture possible sudden changes related to the operation (an infection may itself be the cause of the operation), the day of the operation is isolated and the last week before the operation is subdivided into two time windows. We achieve state-of-the-art accuracy in predicting the development of an infection after surgery from the blood concentrations of fourteen molecules over the four weeks preceding the surgery. Beyond the reliability of the predictions, an analysis reveals that the importance of the features matches the clinicians' intuition, reinforcing the significance of our contribution.

The structure of the rest of the paper is as follows: we begin with a section describing the data set and the proposed features. Section III presents results and analysis followed by a conclusion in Section IV.

A. Related work

Earlier prediction models for surgical site infection rely on statistical analyses of patient and surgery characteristics to

[1] Department of Physics and Technology, UiT The Arctic University of Norway. Contact: jonas.n.myhre@uit.no
[2] Department of Clinical Medicine, UiT The Arctic University of Norway
[3] University Hospital of Northern Norway
[4] NORCE Norwegian Research Centre
Authors AB, JNM, KM, RJ, are all with the UiT Machine Learning Group: http://machine-learning.uit.no

This work was financially supported by Eureka Eurostars project 113519 PERISCOPE, the Research Council of Norway (RCN) through IKTPLUSS grant no. 303514, and the UiT Thematic Initiatives "Data-Driven Health Technology" and "Consortium for Patient-Centered AI.

isolate significant factors. The landmark work of Malone *et al.* [22] demonstrated that the main risk factors of an SSI are diabetes and malnutrition. They define the latter as a weight-loss $> 10\%$ over the six months preceding the surgery. Our data-driven approach is consistent with this result. However, we consider solely preoperative blood tests over the four preceding weeks. Interestingly, Saunders *et al.* [23] recently showed that the hospital ward is also a determining factor.

Baldominos *et al.* [24] perform an extensive review of 101 works published between 2003 and 2020 aiming at predicting infection using computational intelligence. Sepsis comes out as the most studied infection, followed by SSI with twelve studies. Most of the latter leverage a combination of demographic data and clinical variables to monitor the patients after surgery, i.e., using postoperative data [9], [25]. Strauman *et al.* [26] develop a recurrent network-based approach to process the same blood test data set as we do but consider only the postoperative part. At last, Kocbek *et al.* [21], with which we also share the data set, remains the only listed work leveraging preoperative data. However, unlike our approach, the model does not discriminate between the infection's depths.

Kocbek *et al.* [21] gathered 252 features from the 60 last days before the day of surgery. They use a combination of time windows, medium, short and long, and calculate mean blood sample values within these. They also check for abnormal values and count the number of tests taken for each blood value. Finally, they test the influence of the price for each blood test on their classifier. Up to our knowledge, this work is the *only* other work so far that focuses solely on preoperative infection prediction. We encourage the reader to consult Table 1 in their paper, which contains a brief description of papers related to SSI prediction until 2019.

Finally, purely data-driven approaches on unstructured data are becoming ubiquitous thanks to the development of advanced machine learning techniques [27], [28], [29]. Extracting information from text using natural language processing algorithms is especially difficult as the models need to capture different semantics within the words [30]. Karhade *et al.* [31] analyzed free-text notes of patients to report a postsurgical infection automatically. They achieve notable performance on a highly imbalanced data set (about 1.1% positive cases in the training and testing sets). Other works explore the use of deep learning models for monitoring wounds based on images. [32], [33]. This work is also part of this line of research, as we choose a fully data-driven perspective on the problem of prediction SSI from preoperative data. Indeed, our features are computed from raw blood test results. Also, to ensure explainability and dissemination of our results, we process them using classifiers already well known within the medical literature.

II. Methods

A. Data description

The data set is a subset of the database of the University Hospital of North Norway (UNN). The latter contains the electronic health records of 7,741 patients that underwent a gastrointestinal surgical procedure in the years 2004–2012. The ethics approval for the present study was obtained from the Data Inspectorate and the Ethics Committee at the UNN [34].

Similarly to earlier studies [8], [9], we define an SSI and its depth according to the International Classification of Diseases (ICD10) and NOMESCO Classification of Surgical Procedures (NCSP) codes related to severe postoperative complications. We consider only cases where an infection occurred within 30 days after surgery [4]. Among the $1,137$ remaining patients, 132 developed a shallow postoperative infection, 101 a deep infection and 904 did not develop an SSI. The patients are represented using the blood concentrations of the fourteen (14) most frequently monitored molecules or cells recorded in their electronic health records: *alanine aminotransferase* (ALAT), *albumin*, *alkaline phosphatase* (ALP), *amylase*, *aspartate aminotransferase* (ASAT), *bilirubin total*, *creatinine*, *C-reactive protein* (CRP), *glucose*, *hemoglobin*, *leukocytes*, *potassium*, *sodium* and *thrombocytes*. Table VI in the appendix reports for each blood test the unit and the nominal levels with respect to gender and age [35].

B. Features and preprocessing

In this section, we present our new set of features. Three preprocessing steps are carried out in order to deal with common issues in medical data: logarithmic standardization reduces the skeweness of the data [36], class imbalance is handled via over and undersampling [37], [38] and finally, a k-nearest neighbors imputation compensates for the missing data [39].

1) Log-standardization: For each blood component, a nominal range bounds its concentration for a person to be considered healthy (see Table VI, Appendix). Blood test values, which are per se non-negative, yield distributions with positive skews. This implies that *abnormal* values above the nominal range spread over an unbounded interval, while those below are lower-bounded by zero. In terms of variance, this means that the variance of the values lower than the nominal mean is upper-bounded whereas not that of the values larger than mean is not. As we shall see, this asymmetry hinders the classification model (see Table III for an ablation study).

As a remedy, we consider the log of the concentrations and apply a standardization to have a uniform variance [36]. That is, the feature associated to the value of a blood test x is $\frac{\log(x)-\mu}{\sigma/2}$, where μ and σ are the center and the size of the nominal range *after* log-transformation, respectively.

After this log-standardization, x follows a standard Gaussian distribution and both sides of the mean spread symmetrically. Figure 1 illustrate the effect of this log-standardization on the distributions of *sodium* (left column) and *ASAT* (right column). The true distribution of the blood concentration of sodium (top left) presents a strong positive skew and is concentrated around its mode, which is lower than the nominal range (gray area). After transformation, the distribution still has a positive skew but it is more spread. In the case of ASAT (top right), the mode is larger than the nominal range and the distribution has a long tail toward higher values. After the transformation (bottom right), the distribution is slightly more concentrated and it has a shorter right-tail.

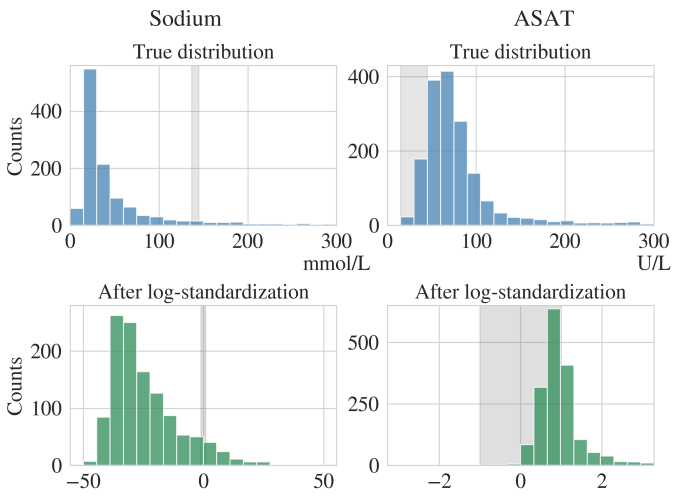

Fig. 1: Histogram of blood sample values for Sodium and ASAT. The nominal range, grey shade, are the corresponding reference values for the blood test, found in Table VI. The log-transformation spreads the distributions on the left of the nominal ranges and contracts them on its right.

We do not standardize the data using the sample mean/mode and standard deviation since we wish to keep as reference a "healthy" individual which is defined by the nominal values. That way, after transformation, positive (negative) values indicate an excess (lack) of a blood component, and larger absolute values mean more severe sickness. Besides, using sample statistics would introduce a bias toward the training set, which would make our approach less general.

2) Time windows: To address the irregularity of data collection, we use averages over six time windows dividing the four weeks preceding the surgery as well as the day of surgery: the day of surgery (day 0), the three days before, the four previous days to complete a week, plus three week long windows from day 8 to 14, 15 to 21 and 22 to 28, respectively. We refer to these as *weekly features*, despite the fact that the last eight recorded days are split into three intervals.

3) Features: The full feature set based on this temporal split consists of 157 features:

- 84 features: the log-transformed average value of fourteen blood tests within each temporal window $(14 \times 6 = 84)$,
- 70 features: the difference between two successive window log-transformed averages, starting from the earliest one $(14 \times 5 = 70)$,
- 2 features: the sex and the age of the patient,
- 1 feature: the logarithm of the number of days since the first recorded blood test for the patient.

Note that the last feature is also logarithmic to account for the positive skew.

In previous studies, [21], the number of blood tests carried out for each patient has shown high correlation with the risk of infection. This is a well known systematic bias: if a medical doctor suspects the patient is ill, more blood tests are ordered [40]. Moreover, this depends on the particular clinicians' discretion and intuition about the patient's condition, therefore we do not include it in our features.

4) Imputation: The lack of a strict data collection protocol results in a sparse data set. For example, on the day of surgery, more than 64% of the data is missing. Two weeks earlier, the percentage exceeds 86%. To account for this, we use a k-nearest neighbours imputation scheme [41], [42], [43]. The missing values are replaced with the average value of the 5-nearest neighbours stratified with respect to the depth of infection and the gender. If a feature is still missing, it is filled with the average of the gender. We include an analysis of the influence of the number of neighbors used in the appendix (Figure 3).

5) Over and undersampling: The data is highly imbalanced : 12.6% and 9.3% of the patients suffered from superficial or deep postoperative infection, respectively. This phenomenon affects both the training of the classification model and its evaluation [44]. As a remedy, the minority classes are oversampled during training using the SMOTE [45] algorithm. However, we undersample the majority class to compute performance metrics on balanced test sets. The majority class is randomly divided into batches of approximately the same size as the minority classes and we report the average metrics over these batches.

C. Classification problems

We consider two classification strategies in this work. The first is a standard one-vs-rest multi-class classification. The other stratifies the task into two binary problems: 1) detect the development of an SSI and, 2) in the positive case, the depth thereof (shallow or deep). A full multi-class classification problem can then be constructed by combining the conditional probabilities. E.g. for the *deep* class this yields:

$$p(\text{depth}) = p(\text{depth}|\text{infection})p(\text{infection}). \quad (1)$$

Contrary to the one-vs-rest scheme, the stratified approach allows for stage-by-stage analysis of the features. That way we compare the rankings of the features for detecting SSIs and for discriminating depths. It is indeed not given that these rankings match.

D. Experimental Setting

1) Predictive models: Two classification algorithms are considered:

- GBOOST: gradient boosted decision trees with logistic loss and Friedman's mean square error as quality measure of a split [46].
- SGD: logistic regression with ℓ_2 regularization and stochastic gradient descent optimization, see e.g. [47] for further details.

The GBOOST model is a widely used classification algorithm that was chosen to enable us to compare results the experimental setting of Kocbek *et al.* [21]. In addition, the GBOOST implementation we used contains a feature importance score, based on the average reduction in impurity score for each feature [48]. Furthermore, we chose the ℓ_2 regularized logistic regression because it is a simple linear model where the feature

TABLE I: Prediction scores for the binary problems of forecasting an SSI (SSI Binary column) or its subtype (Depth Binary column) and the multi-class problem as described in (1) (Multi-class column). For each column, the best results (not statistically different ($p > 0.5$) from the highest score) are marked in bold.

		SSI Binary		Depth Binary		Multi-class	
Features		AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
Kocbek	GBOOST	0.926 ± 0.003	0.960 ± 0.002	0.650 ± 0.023	0.699 ± 0.018	0.776 ± 0.005	0.854 ± 0.003
	SGD	0.922 ± 0.003	0.958 ± 0.002	0.658 ± 0.024	0.688 ± 0.021	0.772 ± 0.024	0.858 ± 0.016
Ours	GBOOST	$\mathbf{0.987 \pm 0.008}$	$\mathbf{0.991 \pm 0.007}$	$\mathbf{0.885 \pm 0.040}$	$\mathbf{0.894 \pm 0.037}$	$\mathbf{0.897 \pm 0.018}$	$\mathbf{0.937 \pm 0.011}$
	SGD	0.984 ± 0.008	0.989 ± 0.007	0.866 ± 0.038	0.881 ± 0.035	0.811 ± 0.028	0.883 ± 0.021

TABLE II: One-versus-rest prediction score for the multi-class problem. For each column, the best results (not statistically different ($p > 0.5$) from the highest score) are marked in bold.

		No SSI-vs-rest		Shallow SSI-vs-rest		Deep SSI-vs-rest	
		AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
Kocbek	GBOOST	0.947 ± 0.003	0.960 ± 0.002	0.705 ± 0.016	0.823 ± 0.007	0.622 ± 0.018	0.820 ± 0.008
	SGD	0.945 ± 0.005	0.958 ± 0.002	0.653 ± 0.025	0.792 ± 0.020	0.632 ± 0.026	0.810 ± 0.021
Ours	GBOOST	$\mathbf{0.985 \pm 0.012}$	$\mathbf{0.991 \pm 0.007}$	$\mathbf{0.869 \pm 0.037}$	$\mathbf{0.919 \pm 0.021}$	$\mathbf{0.881 \pm 0.038}$	$\mathbf{0.940 \pm 0.019}$
	SGD	0.956 ± 0.019	0.967 ± 0.015	0.793 ± 0.047	0.861 ± 0.030	0.821 ± 0.041	0.903 ± 0.026

importance can be directly interpreted through the coefficients of the predictor. We use a 90/10 training-test split and the labels indicate the depth of the SSIs. Each experiment is repeated 100 times and we report the resulting means and standard deviations. For the SGD model, we further split the training-set into several train/validation sets using a 80/20 holdout scheme to select the regularization weight α. The possible values ranges in the $\log_{10}$-space from 0 to -4 with steps of 0.5.

We use the the python implementations of *scikit-learn*[1]. Code to reproduce results can be found in our GitHub repository[2].

2) Performance metrics: To evaluate the predictions, we use the *area under the precision recall curve* (AUPRC) and the *area under receiver operating characteristics* (AUROC). For the multi-class problem, these scores are computed in an one-vs-rest manner. Due to sensitivity of the metrics to the imbalance of the data set, the majority class of the test set is undersampled and we report the average values of the metrics.

We also report specificity, sensitivity, positive predictive value (PPV) and negative predictive value (NPV) for both binary problems in Table V in the appendix.

III. Results and discussion

The task at hand is to predict postoperative infection given preoperative blood tests, and to subsequently predict the infection subtype given a positive infection prediction. We compare our weekly average features to the features devised by Kocbek et al. [21] as they are the current state-of-the-art on the data set . In both cases, the features are standardized before classification.

A. Predicting surgical site infection after surgery: binary and multi-class case

We used the original implementation of Kocbek *et al.* to compute their features, but with a different implementation of GBOOST and hyper-parameters as our implementation is in Python instead of R. Also, we do not use a fixed test set in our setup[3] and we account for class imbalance using oversampling. Table I summarizes classification performances for both features and prediction models.

The results of GBOOST in our implementation on Kocbek's features are slightly different compared to their implementation, but at their advantage. The original paper reports an AUROC of 0.954 for the SSI prediction classification, while our implementation returns 0.960.

Overall, the proposed features yield the highest classification scores, particularly when processed with GBOOST. Our features improve the performance when compared against Kocbek's even with SGD, especially on the binary problem of predicting SSI depths. Regarding Kocbek's features, SGD returns results on par with GBOOST: The difference stays below 0.01 points between SGD and GBOOST in all cases except for the AUROC on the depth binary problem.

B. Predicting either class-versus-rest

Table II details the performances for each class in a one-vs-rest setting. Again, our features processed with GBOOST yield the highest classification scores. The improvements on the classifications of both models is especially noticeable on the deep SSI-vs-rest problem: SGD outperforms GBOOST when using Kocbek's features (AUPRC of 0.632), but it is far from being competitive against SGD and GBOOST when using our features (AUPRC of 0.821 and 0.881 respectively).

C. Preprocessing ablation study

The usage of our weekly features involves two preprocessing steps: the log-standardization and the oversampling. To assess the importance of both steps, we report the prediction performance of GBOOST , using our features without one or either steps. The results are shown in Table III. Note that a

[1]https://scikit-learn.org/stable/index.html

[2]https://github.com/uitml/ssi_prediction

[3]Kocbek *et al.* [21] used a fixed test set for all experiments, based on the AMIA competition split [8].

TABLE III: Prediction scores for GBOOST on our weekly features with (✓) or without (−) log-transformation or oversampling.

		SSI Binary		Depth Binary		Multi-class	
Log-transfo.	Oversampling	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
-	-	0.953 ±0.017	0.965 ±0.015	0.706 ±0.049	0.716 ±0.058	0.980 ±0.009	0.850 ±0.019
✓	-	0.960 ±0.016	0.970 ±0.015	0.705 ±0.051	0.713 ±0.057	0.983 ±0.008	0.853 ±0.018
-	✓	**0.985 ±0.010**	0.989 ±0.008	0.852 ±0.042	0.869 ±0.041	0.994 ±0.005	0.919 ±0.012
✓	✓	**0.987 ±0.008**	**0.991 ±0.007**	**0.885 ±0.040**	**0.894 ±0.037**	**0.995 ±0.004**	**0.937 ±0.011**

standardization using the training set is still performed before any computation.

The use of oversampling alone yields results on the binary prediction of an SSI that are not statistically different from the full model. This is therefore an important step, which is natural since there is quite a high fraction of missing data. The relevance of the logarithmic transformation becomes apparent when the depth of infection is involved. The improved performance suggest that, by acting on the positive skew of the distributions, the logarithmic transformation allows a better separation of patients associated with a deep and superficial infections.

D. Feature ablation study

To evaluate the effect of reduced number of features, we re-trained the GBOOST with our features on both binary problems, ordered the features by importance and removed them one by one starting from the lowest weighted features. In a nutshell, the first setup of this study consists of the GBOOST trained with all features but the least important one, whereas in the last setup only one feature – the most important one – is used. The experiment is repeated 10 times and Figure 2 depicts the evolution of the mean AUPRC and AUROC on both binary problems.

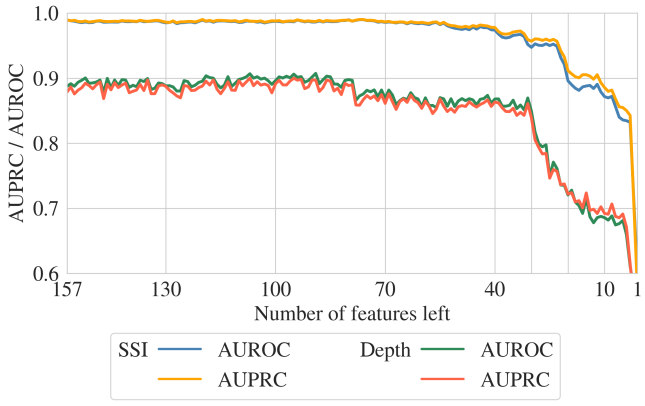

Fig. 2: Prediction scores as a function of the number of features included, from the far left where all the features are included to the far right where only the most important feature is used.

The curves reveal the robustness of both the model and the features. With 20 features left, the AUROC is above 0.9 and 0.7 for the SSI and depth problem, respectively. These results are comparable to the ones obtained by using all the 254 Kocbek's features (see Table II).

The relative stability until 40 and 80 features for the SSI and depth prediction task, respectively, might indicate that most of the features are superfluous or correlated. These plateaus could suggest that less blood samples measurements are needed and more simple protocols could be devised.

E. Feature importance

In Table IV, we report the ten most important weekly features for both binary problems. The ranks are computed as the average rank over the 100 runs previously used in Table I.

Both rankings are topped by features related to the nutritional status (glucose, albumin), liver function (ASAT, ALAT, ALP) and the presence of an infection (trombocytes, CRP). This is not surprising, as the data set comes from a gastrointestinal department.

The results are especially remarkable as they echo several previously published clinical studies. For example, it has been shown that increased levels of sodium before surgery is associated with an increased hospital mortality [49]. A similar conclusion can be drawn here based on the features for the SSI binary problem (Table IV top). The concentration of sodium on day 0 is the fourth most important feature. Moreover, this feature also strongly correlates with the level of potassium between days 4 and 7 prior to surgery (Pearson-$\rho = 0.717$ and p-values < 0.01), which also tops the ranking.

Malnutrition and diabetes are known important risk factors [22], [50], [51], [52], [53] and several related molecules are among the top-10 features for predicting the severity of the SSI, e.g. diabetes and glucose levels (see Table IV bottom). Albumin appears third among these: Bozzetti *et al.* [54] showed that low concentrations of albumin, reflecting a state of malnutrition, correlates with a higher risk of complicated SSI.

It is important to note that, unlike a clinical study, our analysis only allows us to hypothesize that nutritional status is a key factor of SSI. It does not in any way prove a causality. On the other hand, the credibility of the hypothesis benefits from the fact that it is inferred by a fully data-driven model. Since the correlation between SSI and nutrition has already been proven, we can build on the predictions of our model. We believe that thanks to the nature of our features, one could quantify malnutrition and define a cutoff beyond which the risk of surgical site infection is *too* high.

IV. CONCLUSIONS

In this paper we presented a framework for detecting surgical site infections after surgery, based on preoperative features exclusively. We showed that weekly averages of blood test

results can be used as features to predict postoperative infections and their depth with state-of-the-art accuracy. Moreover, the analysis of the features' importance revealed the clinical soundness of our predictions which we believe is essential for integration into clinical decision *support* systems.

The work presented in this paper suggests several potential directions of further research, which require an extended and more regular data collection protocol. Our weekly features can also be interpreted as moving averages of a time series. We consider therefore to investigate several alternative approaches specific to time series. However, to be fully effective and relevant, these methods would require a finer time resolution, i.e., less missing data. Furthermore, electronic health records could provide extra information to stratify based on the condition of the patient or the type of surgery. We would expect this to highlight features specific to each case, leading to a more personalized follow-up of the patients. Finally, robustness and confidence are very important elements in clinical – and general safety oriented – applications. An algorithm should yield similar results for similar patient types or conditions, and also indicate its confidence of the result. Also, if a completely new observation is presented, e.g. a previously unseen condition, the algorithm should mark its prediction as uncertain, preferably indicating *why* it is so. This is currently not possible with standard classification algorithms.

V. Acknowledgment

We would like to thank Paul Costille for the valuable discussions.

TABLE IV: Feature importance for GBoost for the SSI Binary and Depth Binary problem.

SSI Binary		
Feature	Days	Avg.Rank
Potassium	4-7	3.68 ± 3.29
ASAT	8-14	6.66 ± 5.6
ALAT	0	6.82 ± 7.28
Sodium	0	8.0 ± 5.88
CRP	4-7	8.05 ± 6.23
ALAT	15-21	9.24 ± 7.15
Glucose	22-28	9.25 ± 6.95
ALP	8-14	11.52 ± 8.17
ALP	4-7	12.58 ± 8.98
Leukocytes	22-28	13.95 ± 10.13
Depth Binary		
Feature	Days	Avg.Rank
Leukocytes	4-7	4.76 ± 4.67
Glucose	22-28	5.49 ± 7.35
Albumin	8-14	8.72 ± 8.12
ALP	1-3	11.76 ± 14.56
Glucose	1-3	13.43 ± 15.11
ALAT	22-28	14.26 ± 12.57
Potassium	1-3	15.14 ± 14.34
ΔCreatinine	1-3	18.07 ± 20.47
Hemoglobin	0	18.43 ± 22.64
ΔBilirubine tot	1-3	21.85 ± 21.36

References

[1] S. S. Lewis, R. W. Moehring, L. F. Chen, D. J. Sexton, and D. J. Anderson, "Assessing the relative burden of hospital-acquired infections in a network of community hospitals," *Infection Control & Hospital Epidemiology*, vol. 34, no. 11, pp. 1229–1230, 2013.

[2] S. S. Magill, W. Hellinger, J. Cohen, R. Kay *et al.*, "Prevalence of healthcare-associated infections in acute care hospitals in Jacksonville, Florida," *Infection Control*, vol. 33, no. 03, pp. 283–291, 2012.

[3] G. de Lissovoy, K. Fraeman, V. Hutchins, D. Murphy, D. Song, and B. B. Vaughn, "Surgical site infection: incidence and impact on hospital utilization and treatment costs," *American Journal of Infection Control*, vol. 37, no. 5, pp. 387–397, 2009.

[4] D. Nepogodiev, J. Martin, B. Biccard, A. Makupe, A. Bhangu, A. Ademuyiwa, A. O. Adisa, M.-L. Aguilera, S. Chakrabortee, J. E. Fitzgerald *et al.*, "Global burden of postoperative death," *The Lancet*, vol. 393, no. 10170, p. 401, 2019.

[5] B. Allegranzi, P. Bischoff, S. de Jonge, N. Z. Kubilay, B. Zayed, S. M. Gomes, M. Abbas, J. J. Atema, S. Gans, M. van Rijen *et al.*, "New who recommendations on preoperative measures for surgical site infection prevention: an evidence-based global perspective," *The Lancet Infectious Diseases*, vol. 16, no. 12, pp. e276–e287, 2016.

[6] C. Y. Ko, B. L. Hall, A. J. Hart, M. E. Cohen, D. B. Hoyt *et al.*, "The American college of surgeons national surgical quality improvement program: achieving better and safer surgery," *The Joint Commission Journal on Quality and Patient Safety*, vol. 41, no. 5, pp. 199–AP1, 2015.

[7] P. L. Owens, M. L. Barrett, S. Raetzman, M. Maggard-Gibbons, and C. A. Steiner, "Surgical site infections following ambulatory surgery procedures," *JAMA*, vol. 311, no. 7, pp. 709–716, 2014.

[8] C. Soguero-Ruiz, R. Jenssen, K. M. Augestad, S. O. Skrøvseth *et al.*, "Data-driven temporal prediction of surgical site infection," in *AMIA Annual Symposium Proceedings*, vol. 2015. American Medical Informatics Association, 2015, p. 1164.

[9] P. C. Sanger, G. H. van Ramshorst, E. Mercan *et al.*, "A prognostic model of surgical site infection using daily clinical wound assessment," *Journal of the American College of Surgeons*, vol. 223, no. 2, pp. 259 – 270.e2, 2016.

[10] E. H. Lawson, B. L. Hall, and C. Y. Ko, "Risk factors for superficial vs deep/organ-space surgical site infections: implications for quality improvement initiatives," *JAMA surgery*, vol. 148, no. 9, pp. 849–858, 2013.

[11] J. Blumetti, M. Luu, G. Sarosi, K. Hartless, J. McFarlin, B. Parker, S. Dineen, S. Huerta, M. Asolati, E. Varela *et al.*, "Surgical site infections after colorectal surgery: do risk factors vary depending on the type of infection considered?" *Surgery*, vol. 142, no. 5, pp. 704–711, 2007.

[12] J. D. Whitehouse, N. D. Friedman, K. B. Kirkland, W. J. Richardson, and D. J. Sexton, "The impact of surgical-site infections following orthopedic surgery at a community hospital and a university hospital adverse quality of life, excess length of stay, and extra cost," *Infection Control & Hospital Epidemiology*, vol. 23, no. 04, pp. 183–189, 2002.

[13] R. Shah, E. Pavey, M. Ju, R. Merkow *et al.*, "Evaluation of readmissions due to surgical site infections: A potential target for quality improvement," *The American Journal of Surgery*, 2017.

[14] J. Silvestre, J. Rebanda, C. Lourenço, and P. Póvoa, "Diagnostic accuracy of C-reactive protein and procalcitonin in the early detection of infection after elective colorectal surgery–a pilot study," *BMC infectious diseases*, vol. 14, no. 1, p. 444, 2014.

[15] F. J. Medina-Fernández, D. J. Garcilazo-Arismendi, R. García-Martín, L. Rodríguez-Ortiz, J. Gómez-Barbadillo *et al.*, "Validation in colorectal procedures of a useful novel approach for the use of C-reactive protein in postoperative infectious complications," *Colorectal Disease*, vol. 18, no. 3, pp. O111–O118, 2016.

[16] C. Soguero-Ruiz, A. Revhaug, R.-O. Lindsetmo, K. M. Augestad, R. Jenssen *et al.*, "Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records," *IEEE journal of biomedical and health informatics*, vol. 20, no. 5, pp. 1404–1415, 2016.

[17] Z. Hu, G. B. Melton, E. G. Arsoniadis, Y. Wang, M. R. Kwaan, and G. J. Simon, "Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record," *Journal of Biomedical Informatics*, vol. 68, pp. 112–120, 2017.

[18] M. R. Angiolini, F. Gavazzi, C. Ridolfi, M. Moro, P. Morelli, M. Montorsi, and A. Zerbi, "Role of C-reactive protein assessment as early predictor of surgical site infections development after pancreaticoduodenectomy," *Digestive surgery*, vol. 33, no. 4, pp. 267–275, 2016.

[19] S. L. Gans, J. J. Atema, S. Van Dieren, B. G. Koerkamp, and M. A. Boermeester, "Diagnostic value of C-reactive protein to rule out infectious complications after major abdominal surgery: a systematic review

and meta-analysis," *International journal of colorectal disease*, vol. 30, no. 7, pp. 861–873, 2015.
[20] E. Mujagic, W. R. Marti, M. Coslovsky, J. Zeindler, S. Staubli, R. Marti, R. Mechera, S. D. Soysal, L. Gürke, and W. P. Weber, "The role of preoperative blood parameters to predict the risk of surgical site infection," *The American Journal of Surgery*, vol. 215, no. 4, pp. 651–657, 2018.
[21] P. Kocbek, N. Fijacko, C. Soguero-Ruiz, K. Ø. Mikalsen, U. Maver, P. Povalej Brzan, A. Stozer, R. Jenssen, S. O. Skrøvseth, and G. Stiglic, "Maximizing interpretability and cost-effectiveness of surgical site infection (ssi) predictive models using feature-specific regularized logistic regression on preoperative temporal data," *Computational and mathematical methods in medicine*, vol. 2019, 2019.
[22] D. L. Malone, T. Genuit, J. K. Tracy, C. Gannon, and L. M. Napolitano, "Surgical site infections: reanalysis of risk factors," *Journal of Surgical Research*, vol. 103, no. 1, pp. 89–95, 2002.
[23] L. Saunders, M. Perennec-Olivier, P. Jarno, F. L'Hériteau, A.-G. Venier, L. Simon, M. Giard, J.-M. Thiolet, J.-F. Viel, and R. group, "Improving prediction of surgical site infection risk with multilevel modeling," *PloS one*, vol. 9, no. 5, p. e95295, 2014.
[24] A. Baldominos, A. Puello, H. Oğul, T. Aşuroğlu, and R. Colomo-Palacios, "Predicting infections using computational intelligence–a systematic review," *IEEE Access*, vol. 8, pp. 31 083–31 102, 2020.
[25] C. Ke, Y. Jin, H. Evans, B. Lober, X. Qian, J. Liu, and S. Huang, "Prognostics of surgical site infections using dynamic health data," *Journal of Biomedical Informatics*, vol. 65, pp. 22–33, 2017.
[26] A. S. Strauman, F. M. Bianchi, K. Ø. Mikalsen, M. Kampffmeyer, C. Soguero-Ruiz, and R. Jenssen, "Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2018, pp. 307–310.
[27] C. F. Luz, M. Vollmer, J. Decruyenaere, M. W. Nijsten, C. Glasner, and B. Sinha, "Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies," *Clinical Microbiology and Infection*, vol. 26, no. 10, pp. 1291–1299, 2020.
[28] S. Al-Ahmari and F. Nadeem, "Machine learning-based predictive model for surgical site infections: A framework," in *2021 National Computing Colleges Conference (NCCC)*. IEEE, 2021, pp. 1–6.
[29] A. Samareh, X. Chang, W. B. Lober, H. L. Evans, Z. Wang, X. Qian, and S. Huang, "Artificial intelligence methods for surgical site infection: impacts on detection, monitoring, and decision making," *Surgical infections*, vol. 20, no. 7, pp. 546–554, 2019.
[30] D. A. da Silva, C. S. Ten Caten, R. P. Dos Santos, F. S. Fogliatto, and J. Hsuan, "Predicting the occurrence of surgical site infections using text mining and machine learning," *PloS one*, vol. 14, no. 12, p. e0226272, 2019.
[31] A. V. Karhade, M. E. Bongers, O. Q. Groot, T. D. Cha, T. P. Doorly, H. A. Fogel, S. H. Hershman, D. G. Tobert, A. J. Schoenfeld, J. D. Kang *et al.*, "Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy?" *The Spine Journal*, vol. 20, no. 10, pp. 1602–1609, 2020.
[32] V. N. Shenoy, E. Foster, L. Aalami, B. Majeed, and O. Aalami, "Deepwound: Automated postoperative wound assessment and surgical site surveillance through convolutional neural networks," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1017–1021.
[33] B. Chang, Z. Sun, P. Peiris, E. S. Huang, E. Benrashid, and E. D. Dillavou, "Deep learning-based risk model for best management of closed groin incisions after vascular surgery," *Journal of Surgical Research*, vol. 254, pp. 408–416, 2020.
[34] K. Jensen, C. Soguero-Ruiz, K. Ø. Mikalsen, R.-O. Lindsetmo, I. Kouskoumvekaki, M. Girolami, S. O. Skrovseth, and K. M. Augestad, "Analysis of free text in electronic health records for identification of cancer patient trajectories," *Scientific Reports*, vol. 7, p. 46226, 2017.
[35] T. Hagve, A. Brun, G. Hov, M. Lindberg, and A. AAsberg, "Nasjonal brukerhåndbok i medisinsk biokjemi," 2014.
[36] M. Bland, *An introduction to medical statistics*. Oxford University Press (UK), 2015.
[37] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing*, vol. 3, no. 2, p. 224, 2013.
[38] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Information Sciences*, vol. 513, pp. 429–441, 2020.
[39] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen, "Missing data and multiple imputation in clinical epidemiological research," *Clinical epidemiology*, vol. 9, p. 157, 2017.
[40] G. Hripcsak and D. J. Albers, "High-fidelity phenotyping: richness and freedom from bias," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 289–294, 2018.
[41] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580–585, 1985.
[42] K.-Y. Kim, B.-J. Kim, and G.-S. Yi, "Reuse of imputed data in microarray analysis increases imputation efficiency," *BMC bioinformatics*, vol. 5, no. 1, pp. 1–9, 2004.
[43] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.
[44] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.
[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
[46] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.
[47] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.
[48] A. Criminisi, J. Shotton, E. Konukoglu *et al.*, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," 2012.
[49] M. Cecconi, H. Hochrieser, M. Chew, M. Grocott, A. Hoeft, A. Hoste, I. Jammer, M. Posch, P. Metnitz, P. Pelosi *et al.*, "Preoperative abnormalities in serum sodium concentrations are associated with higher in-hospital mortality in patients undergoing major surgery," *BJA: British Journal of Anaesthesia*, vol. 116, no. 1, pp. 63–69, 2016.
[50] S. Yamashita, H. Yamaguchi, M. Sakaguchi, T. Satsumae, S. Yamamoto, and F. Shinya, "Longer-term diabetic patients have a more frequent incidence of nosocomial infections after elective gastrectomy," *Anesthesia & Analgesia*, vol. 91, no. 5, pp. 1176–1181, 2000.
[51] D. L. Waitzberg, H. Saito, L. D. Plank, G. G. Jamieson, P. Jagannath, T.-L. Hwang, J. M. Mijares, and D. Bihari, "Postsurgical infections are reduced with specialized nutrition support," *World journal of surgery*, vol. 30, no. 8, pp. 1592–1604, 2006.
[52] C. A. Oh, D. H. Kim, S. J. Oh, M. G. Choi, J. H. Noh, T. S. Sohn, J. M. Bae, and S. Kim, "Nutritional risk index as a predictor of postoperative wound complications after gastrectomy," *World Journal of Gastroenterology: WJG*, vol. 18, no. 7, p. 673, 2012.
[53] Y. Fukuda, K. Yamamoto, M. Hirao, K. Nishikawa, S. Maeda, N. Haraguchi, M. Miyake, N. Hama, A. Miyamoto, M. Ikeda *et al.*, "Prevalence of malnutrition among gastric cancer patients undergoing gastrectomy and optimal preoperative nutritional support for preventing surgical site infections," *Annals of surgical oncology*, vol. 22, no. 3, pp. 778–785, 2015.
[54] F. Bozzetti, L. Gianotti, M. Braga, V. Di Carlo, and L. Mariani, "Postoperative complications in gastrointestinal cancer patients: the joint role of the nutritional status and the nutritional support," *Clinical Nutrition*, vol. 26, no. 6, pp. 698–709, 2007.
[55] C. Bennett and T. Doub, "Data mining and electronic health records: Selecting optimal clinical treatments in practice," in *In: Proc. of the 6th IC on Data Mining*, Las Vegas, Nevada, USA, 2010, pp. 313–18.

APPENDIX

Table V extends Table I and Table II with four additional metrics, commonly found in the literature: specificity, sensitivity, positive predictive value (PPV) and negative predictive value (NPV).

Table VI shows the nominal values for the blood tests used in this study.

Using the experimental setting of Section III-D, we study here the influence of the number of neighbors used for imputation (Section II-B4). Figure 3 depicts the evolution of the mean AUPRC and AUROC for both binary problems. A mean imputation ($k = 0$) performs poorly. In case of kNN, the performance sightly increases until $k = 9$ after which it drops.

Our choice of $k = 5$ lies in the middle. It also represents a trade-off between too few and too much neighbor information, both cases inducing their own bias [42], [55].

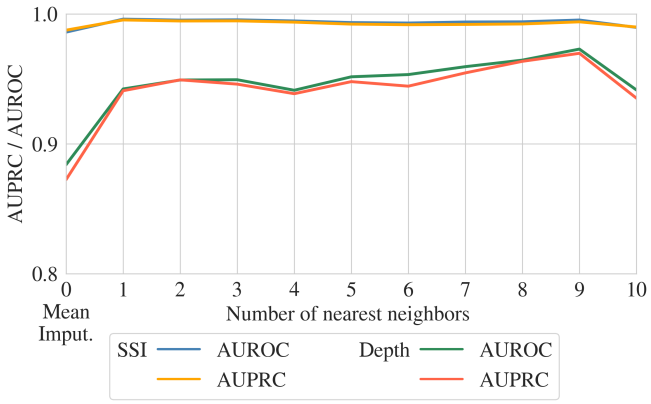

Fig. 3: Prediction scores as a function of the number of neighbors k used for imputation. The case $k = 0$ corresponds to a mean imputation.

TABLE V: Extra prediction scores for the binary problems of forecasting an SSI, its subtype and the three one-versus-rest binary problems.

		SSI Binary					
		AUPRC	AUROC	Sensitivity	Specificity	PPV	NPV
Kocbek	GBoost	0.926 ± 0.003	0.960 ± 0.002	0.737 ± 0.009	0.967 ± 0.002	0.924 ± 0.004	0.879 ± 0.003
	SGD	0.922 ± 0.003	0.958 ± 0.002	0.924 ± 0.007	0.848 ± 0.007	0.764 ± 0.008	0.957 ± 0.004
Ours	GBoost	0.987 ± 0.008	0.991 ± 0.007	0.874 ± 0.026	0.987 ± 0.013	0.979 ± 0.016	0.934 ± 0.013
	SGD	0.984 ± 0.008	0.989 ± 0.007	0.748 ± 0.061	0.984 ± 0.027	0.982 ± 0.015	0.879 ± 0.023
		Depth Binary					
		AUPRC	AUROC	Sensitivity	Specificity	PPV	NPV
Kocbek	GBoost	0.650 ± 0.023	0.699 ± 0.018	0.497 ± 0.027	0.797 ± 0.020	0.662 ± 0.032	0.679 ± 0.013
	SGD	0.658 ± 0.024	0.688 ± 0.021	0.588 ± 0.062	0.633 ± 0.074	0.579 ± 0.037	0.685 ± 0.030
Ours	GBoost	0.885 ± 0.040	0.894 ± 0.037	0.744 ± 0.069	0.858 ± 0.052	0.836 ± 0.056	0.835 ± 0.038
	SGD	0.866 ± 0.038	0.881 ± 0.035	0.744 ± 0.062	0.838 ± 0.059	0.814 ± 0.051	0.830 ± 0.035
		No SSI-vs-rest					
		AUPRC	AUROC	Sensitivity	Specificity	PPV	NPV
Kocbek	GBoost	0.947 ± 0.003	0.960 ± 0.002	0.967 ± 0.002	0.737 ± 0.009	0.711 ± 0.007	0.973 ± 0.002
	SGD	0.945 ± 0.005	0.958 ± 0.002	0.848 ± 0.007	0.924 ± 0.007	0.884 ± 0.010	0.907 ± 0.004
Ours	GBoost	0.985 ± 0.012	0.991 ± 0.007	0.987 ± 0.013	0.874 ± 0.026	0.836 ± 0.029	0.992 ± 0.007
	SGD	0.956 ± 0.019	0.967 ± 0.015	0.984 ± 0.027	0.748 ± 0.061	0.727 ± 0.041	0.993 ± 0.008
		Shallow SSI-vs-rest					
		AUPRC	AUROC	Sensitivity	Specificity	PPV	NPV
Kocbek	GBoost	0.705 ± 0.016	0.823 ± 0.007	0.521 ± 0.019	0.876 ± 0.011	0.698 ± 0.024	0.780 ± 0.008
	SGD	0.653 ± 0.025	0.792 ± 0.020	0.454 ± 0.063	0.863 ± 0.025	0.669 ± 0.041	0.759 ± 0.018
Ours	GBoost	0.869 ± 0.037	0.919 ± 0.021	0.644 ± 0.065	0.923 ± 0.027	0.842 ± 0.050	0.832 ± 0.026
	SGD	0.793 ± 0.047	0.861 ± 0.030	0.556 ± 0.069	0.922 ± 0.025	0.820 ± 0.061	0.798 ± 0.025
		Deep SSI-vs-rest					
		AUPRC	AUROC	Sensitivity	Specificity	PPV	NPV
Kocbek	GBoost	0.622 ± 0.018	0.820 ± 0.008	0.437 ± 0.026	0.921 ± 0.008	0.675 ± 0.034	0.823 ± 0.007
	SGD	0.632 ± 0.026	0.810 ± 0.021	0.517 ± 0.043	0.859 ± 0.022	0.587 ± 0.033	0.837 ± 0.010
Ours	GBoost	0.881 ± 0.038	0.940 ± 0.019	0.674 ± 0.065	0.953 ± 0.020	0.870 ± 0.051	0.892 ± 0.019
	SGD	0.821 ± 0.041	0.903 ± 0.026	0.604 ± 0.082	0.949 ± 0.022	0.843 ± 0.051	0.871 ± 0.023

TABLE VI: Nominal levels for different blood tests found on brukerhandboken.no

Blood test	Unit	Age	Reference level	
			Female	Male
ALAT	U/L	-	10-45	10-70
		18-39	36-48	36-48
Albumin	g/L	40-69	36-45	36-45
		> 69	34-45	34-45
ALP	U/L	-	35-400	35-400
Amylase	U/L	-	25–120	25-120
ASAT	U/L	-	15-35	15-45
Bilirubine total	μmol/L	> 18	5-25	5-25
Creatinine	μmol/L	-	45-90	60-105
CRP	mg/L	-	<4	<4
Glucose (Serum)	mmol/L	-	4–6	4–6
Hemoglobin B	g/dL	-	11.7-15.3	13.4-17.0
Leukocytes	10^9/L	-	4-11	4-11
Potassium	mmol/L	-	3.5-4.4	3.5-4.4
Sodium	mmol/L	-	137-145	137-145
Trombocytes	10^9/L	-	145-390	145-390