2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871788

Using Operative Reports to Predict Heart Transplantation Survival*

Marcus Klang[1], Daniel Diaz[2], Dennis Medved[3], Pierre Nugues[4], and Johan Nilsson[5]

Abstract— **Heart transplantation is a difficult procedure compared with other surgical operations, with a greater outcome uncertainty such as late rejection and death. We can model the success of heart transplants from predicting factors such as the age, sex, diagnosis, etc., of the donor and recipient. Although predictions can mitigate the uncertainty on the transplantation outcome, their accuracy is far from perfect. In this paper, we describe a new method to predict the outcome of a transplantation from textual operative reports instead of traditional tabular data. We carried out an experiment on 300 surgical reports to determine the survival rates at one year and five years. Using a truncated TF-IDF vectorization of the texts and logistic regression, we could reach a macro F1 of 59.1%, respectively, 54.9% with a five-fold cross validation. While the size of the corpus is relatively small, our experiments show that the operative textual sources can discriminate the transplantation outcomes and could be a valuable additional input to existing prediction systems.**

***Clinical relevance*—Heart transplantation involves a significant number of written reports including in the preoperative examinations and operative documentation. In this paper, we show that these written reports can predict the outcome of the transplantation at one and five years with macro F1s of 59.1% and 54.9%, respectively and complement existing prediction methods.**

I. Introduction

Heart transplantation has enabled the survival of patients with advanced or terminal heart diseases. While now well-mastered with thousands of operations performed annually worldwide, heart transplants are still heavy operations with relatively uncertain outcomes compared with other more routine operations. As of today, the 1-year survival is of 91% and the median survival is of 12 to 13 years [5], with relatively important variations across the transplantation sites, while the 10-year survival rate is of 71% in Sweden [7].

In addition to being a complex surgery operation, heart transplantation also involves significant preoperative, care-intensive, and follow-up treatments. In contrast to milder diseases such as, for instance, seasonal influenza, a heart transplantation is documented by an important number of medical analyses and reports. The aim of all these procedures is to reduce the initial uncertainty on the patient survival and mitigate the risks with a personalized follow-up [3].

Nonetheless, while essential to the treatment, the data collected from the donor, the patient, and the operation are sometimes difficult to bring together, even for specialists. Algorithms can help in the decision process, as for instance to assess the compatibility of an organ and a patient [3] or to predict the survival rate from characteristics from the donor and the recipient [10, 9].

To the best of our knowledge, in heart transplantation, these decision support algorithms only use numerical or categorical data as input. They then ignore the textual reports as data sources although these reports form an important component of the medical analyses and a significant information source in the manual determination of the treatment procedure.

In this paper, we describe a corpus of preoperative and operative reports and how we used them to predict the survival outcome of heart transplants. We show that text is useful in the prediction of survival rates at one and five years with macro F1s of 59.1% and 54.9%, respectively. In addition, we determined the most predictive words and we extracted them from the reports, paving the way to outline the most relevant parts of a text.

The Ethics Committee for Clinical Research at Lund University, Sweden, approved the study protocol. The data was anonymized and de-identified prior to analysis and the institutional review board waived the need for written informed consent from the participants.

II. Previous Work

Previous studies on the prediction of transplantation outcomes for different organs include liver, using a variety of methods, such as logistic regression, multilayer perceptron, and transformers [11], kidney, using Cox regression [13], as well as heart transplantation, using logistic regression, neural networks, or deep learning techniques [15, 10, 9]. As predictors, most studies used biological data in a numerical or categorical format.

In this paper, we considered the text of operative reports as input and the survival of the patient one year and five years after transplantation as output. We can then frame the outcome prediction as a text categorization problem: Whether the patient has survived or not one year, respectively five years, after her/his operation.

Text categorization has been applied in many applications, including spam detection, sentiment analysis, movie or

*This work was partially supported by the Heart Lung Foundation, registration number 2019-0623 and *Vetenskaprådet*, the Swedish Research Council, registration number 2021-04533.

[1]Marcus Klang is with the Department of Computer Science, Lund University, Lund, Sweden marcus.klang@cs.lth.se

[2]Daniel Diaz did this work while at the Department of Computer Science, Lund University, Lund, Sweden daniel.diaz.quilez@alumnos.upm.es

[3]Dennis Medved is with the Medicine Faculty, Lund University, Lund, Sweden dennis.medved@med.lu.se

[4]Pierre Nugues is with the Department of Computer Science, Lund University, Lund, Sweden pierre.nugues@cs.lth.se

[5]Johan Nilsson is with the Department of Translational Medicine, Thoracic Surgery and Bioinformatics, Lund University and Skåne University Hospital, Lund, Sweden johan.nilsson@med.lu.se

product reviews with a wide array of techniques including logistic regression, support vector machines [8], multilayer perceptron, recurrent neural networks, and transformers [4].

In the field of surgical operations, only a few papers include predictors in the form of textual descriptions, for instance to predict the duration of a variety of pediatric operations [6]. For transplantation, the references are even sparser, and we could find only one paper by Placona et al. [12] on the prediction of the suitability of an organ in kidney transplants. These authors used the text extracted from donor records, notably the admission course history, donor medical and social history, and modeled the problem as a binary classification with two possible outcomes for the organ: *placed* or *difficult to place*. They vectorized the texts with the TF-IDF method and classified them with logistic regression and a ridge regularization.

To the best of our efforts, we could not find studies using operative reports to predict heart transplantation survival from text. In addition, contrary to [12], we focused on the analysis of recipients instead of organ donors.

III. Materials and Methods

A. Corpus

As corpus, we used a collection of 300 operative reports of living and dead patients, with all of the reports annotated with the date of transplantation and, in case of death of the patient, the date of it. We assigned the negative class to patients who did not survive one year and the positive one to patients who survived more than one or five years.

All of the texts are in Swedish and consist of two sections on the preoperative evaluation and on operative descriptions. The text below shows an except of an original report. It contains a few spelling mistakes and abbreviations:

> **Preop bedömning.** Dilaterad kardiomyopati med progressiv hjärtsvikt som föranleder utredning för hjärtx...
>
> **Operation.** Median resternotomi och kommer in i pericardiet, hjärtat välskyddat av membran likaså utflödesgraftet...

and its translation in English:

> '**Preoperative evaluation.** Dilated cardiomyopathy with progressive heart failure that justifies an investigation for heart transplantation...'
>
> '**Operation.** Median sternotomy and enters the pericardium, the heart well protected by membranes as well as the outflow graft...'

The corpus has about 101,000 words in total with 7,926 unique lowercased words. Out of these unique words, 4,144 occur only once in the corpus (*hapax legomenon*). Figure 1 shows the number of lowercased words by frequency, starting with words occurring once, twice, etc. The five most frequent words are, by far, grammatical words: *och*, 'and', *med*, 'with', *i*, 'in', *på*, 'on', and *av*, 'of'. In this list, *hjärtat* 'heart', is the highest ranked word, not being a stopword.

The text lengths range from 72 to 1,491 words with a mean of about 340 words. Figure 2 shows the distribution of text lengths.

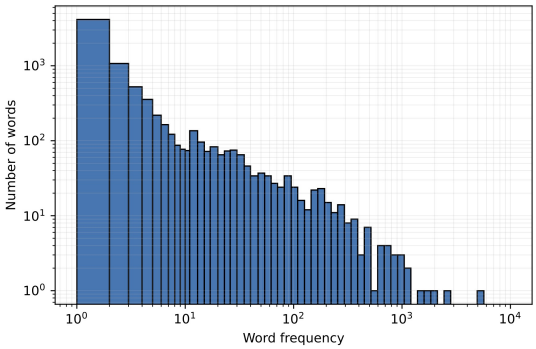

Fig. 1. Distribution of word frequencies (semi-logarithmic scales)

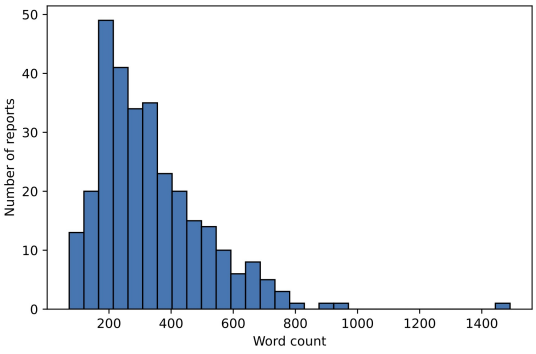

Fig. 2. Distribution of text lengths in the corpus

Out of the 300 patients in the dataset, 98 died, 22 within one year, and 48 within five years. For the one-year survival rate, this corresponds to a highly biased split between the positive and negative classes of 92.7% and 7.3%, and for the five-year one to 83.7% and 16.3%.

B. Algorithms

We evaluated three algorithms:

1) Logistic regression with, respectively: A one-hot encoding of the words; a TF-IDF vectorization; and a dimension reduction that we applied to the vectors with a singular value decomposition (SVD);
2) A neural network with the same vectorization methods;
3) A multilingual transformer model, mBERT.

1) Logistic regression with a bag of words: As input layer, we tried first a bag-of-word vectorization, where each dimensions correspond to one word in the corpus. The length of a vector representing a document is then the number of different words in the corpus. We tried two configurations. The first one with a one-hot encoding and a the second one with parameters being the word frequency in the document.

TF-IDF is an enhanced version of the bag of words. This technique reduces the weight of a word in proportion to the number of reports that contains it. Ultimately, the weight of a word that shows in every text is zero since it does not provide any information.

2) Neural networks: The neural network architecture is a feedforward neural network. It consists of one hidden layer with 16 nodes and a ReLU activation, and an output layer with a single node that uses a sigmoid activation. We used the same vectorization techniques as for logistic regression.

3) Transformers: Transformers are new architectures intended to model long-range dependencies. Transformers stem from more general encoder-decoder architectures, firstly introduced for translation. The encoder vectorizes plain text to represent the contextualized meaning of words, while the decoder translates the vectors into a given language. We used a transformer restricted to the encoder part: multilingual BERT (mBERT) [4]. mBERT is pretrained on a large multilingual corpus with a masked language model, where it learns to predict words missing from a sentence.

Once trained on a large corpus, we used mBERT as an embedding layer. We applied the mBERT tokenizer and the pretrained model to each input document. For each word from an input document, mBERT outputs a 768 long vector. We used the pooled output of these vectors, which is provided by the model, to serve as input to a logistic regression layer. We could then train a model from the mBERT embeddings.

IV. Experimental Setup

As tools, we used the logistic regression and multilayer perceptron modules from scikit-learn and the mBERT model from HuggingFace. For the logistic regression classifier and multilayer perceptron, we evaluated different input configurations as described in Sect. III-B.1 with the original vectors and with a SVD to reduce the number of dimensions to 44 for the one-year survival rate prediction and to 64 for the five-year survival rate.

As the dataset is imbalanced, we used the SMOTE algorithm [1] to oversample the dead patients with a balance ratio of 0.75. We trained the different systems on 300 texts with a stratified 5-fold cross validation with approximately 240 samples for the training set and 60 for the test set (before oversampling).

V. Results

Training the classifier on the heart transplant dataset proved rather unstable because of its small size. We repeated the training process 20 times for each configuration to smooth the scores. This method gave us better estimations, but a bigger dataset would be welcome to reach more robust conclusions.

We also computed the area under the receiver operating characteristic (AUROC) to compare our figures with those of previous works using tabular data such as in [16] and [9]. Finally, we included the Matthews correlation coefficient (MCC) to measure the quality of the classification [2].

Tables I and II show the results we obtained. We also included the 95% confidence interval. The first row shows the scores we obtained with a majority baseline, when we always select the majority class (survive). We only report one vectorization method for the multilayer perceptron.

For the one-year survival, the truncated vectorizations proved better on the macro F1, reaching 59.1% for TF-IDF, showing they can probably discard noisy or useless dimensions in the truncation. The MCC figures are also all positive showing a correlation between the predictions and the facts. Finally, the AUROC reached 0.683 with a bag-of-word vector of counts and proved better than that from previously reported experiments [16, 9] trained on tabular data. These figures are not comparable though as our dataset is smaller and the validation cohort is not identical.

The results on the five-year survival predictions are somewhat lower than for the one-year ones. This was expected as it is always more difficult to predict longer-term outcomes. As for the one-year survival rates, the truncated TF-IDF obtains the best F1 scores with 54.9% with a positive MCC. This method also ranks well with the AUROC score, but lower than a simpler binary bag-of-words encoding.

Finally, in our experiments, mBERT reached scores lower than the other methods, possibly due to the small size of our corpus. A possible future investigation would be to fine-tune the last layers of the mBERT model with the corpus of operative reports.

VI. Interpretation

We used the LIME technique [14] to extract the words significant in the classification process. Table III shows the six most frequent words contributing to the negative and positive class predictions for a one-year survival. Although certain words are difficult to interpret, some of them give valuable clues on the perception of the situation by the surgeon:

- For the negative class, the patients not surviving, ECMO is ranked first and stands for *extracorporeal membrane in oxygenation*. It is a device to support the circulatory function indicating a very severe heart failure. It reflects the patient's severe condition hinting at a possible imminent death. *Anastomosis* is a connection between vessels, which could indicate technical issues during the surgery.
- For the positive class, LVAD and HeartMate are assisting devices carried by a patient before the operation, indicating a better physical condition before the transplantation. LVAD is a generic name, while HeartMate is a specific trademark.

VII. Conclusion

In this paper, we showed that predictions of one-year and five-year survival rates can be made from plain text descriptions matching or surpassing methods with tabular data. We evaluated different architectures, where we found that a text vectorization consisting of a TF-IDF bag of words followed by a truncation and a logistic regression classifier is a solid choice, at least when dealing with a small dataset. The results we obtained would benefit nonetheless from a confirmation on a larger patient cohort.

As future work, we plan to collect more reports as well as to include tabular data in our models. We hope applying

TABLE I

RESULTS FOR THE PREDICTION OF THE ONE-YEAR SURVIVAL RATE

Input vector	**Classifier**	**F1 score**	**MCC**	**AUROC**
Baseline	Majority	0.481	0.0	0.5
Bag of words (binary)	LR	0.488 ± 0.008	0.031 ± 0.035	0.626 ± 0.015
Bag of words (counts)	LR	0.538 ± 0.009	**0.206** ± 0.025	**0.683** ± 0.016
Bag of words (truncated)	LR	0.535 ± 0.015	0.073 ± 0.033	0.664 ± 0.027
TF-IDF (truncated)	LR	**0.591** ± 0.017	0.189 ± 0.034	0.680 ± 0.016
TF-IDF (truncated)	MLP	0.579 ± 0.020	0.167 ± 0.041	0.663 ± 0.021
mBERT	LR	0.506 ± 0.008	0.015 ± 0.016	0.475 ± 0.017

TABLE II

RESULTS FOR THE PREDICTION OF THE FIVE-YEAR SURVIVAL RATE

Input vector	**Classifier**	**F1 score**	**MCC**	**AUROC**
Baseline	Majority	0.457	0.0	0.5
Bag of words (binary)	LR	0.482 ± 0.007	0.073 ± 0.025	**0.600** ± 0.015
Bag of words (counts)	LR	0.502 ± 0.011	0.063 ± 0.030	0.550 ± 0.012
Bag of words (truncated)	LR	0.497 ± 0.013	0.004 ± 0.027	0.524 ± 0.013
TF-IDF (truncated)	LR	**0.549** ± 0.013	**0.102** ± 0.026	0.576 ± 0.011
TF-IDF (truncated)	MLP	0.536 ± 0.010	0.077 ± 0.020	0.566 ± 0.014
mBERT	LR	0.459 ± 0.007	-0.050 ± 0.016	0.465 ± 0.014

TABLE III

TOP PREDICTING WORDS FOR ONE-YEAR SURVIVAL WITH A TRUNCATED TF-IDF VECTORIZATION AND LOGISTIC REGRESSION

Negative class	Counts	Positive class	Counts
ECMO	17	HeartMate	23
patienten 'patient'	15	Prolene	11
bicavala 'cava venae canulation'	8	LVAD	8
ortotop 'heart transplantation'	7	härefter 'hence'	6
donatorshjärta 'donated heart'	5	Prolen	5
anastomosen 'anastomosis'	4	driveline	3

this kind of prediction to operative reports of transplanted patients will help improve their personalized follow-up treatment.

REFERENCES

[1] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[2] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1 (2020), pp. 1–13.

[3] Maria Rosa Costanzo et al. "The International Society of Heart and Lung Transplantation Guidelines for the care of heart transplant recipients". In: *The Journal of heart and lung transplantation* 29.8 (2010), pp. 914–956.

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[5] Eileen M. Hsich et al. "Heart Transplantation: An In-Depth Survival Analysis". In: *JACC: Heart Failure* 8.7 (2020), pp. 557–568.

[6] York Jiao et al. "Probabilistic forecasting of surgical case duration using machine learning: model development and validation". In: *Journal of the American Medical Informatics Association* 27.12 (2020), pp. 1885–1893.

[7] B. Kornhall et al. "Fler hjärttransplantationer än någonsin 'More heart transplantations than ever'". In: *Läkartidningen* 109.39-40 (2012), pp. 1743–1744.

[8] David D. Lewis et al. "RCV1: A New Benchmark Collection for Text Categorization Research". In: *Journal of Machine Learning Research* 5 (2004), pp. 361–397.

[9] Dennis Medved et al. "Improving prediction of heart transplantation outcome using deep learning techniques". In: *Scientifc Reports* 8.3613 (2018).

[10] Johan Nilsson et al. "The International Heart Transplant Survival Algorithm (IHTSA): a new model to improve organ sharing and survival". In: *PloS one* 10.3 (2015), e0118644.

[11] Osvald Nitski et al. "Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data". In: *The Lancet Digital Health* 3.5 (2021), e295–e305.

[12] Andrew M Placona et al. "Can donor narratives yield insights? A natural language processing proof of concept to facilitate kidney allocation". In: *American Journal of Transplantation* 20.4 (2020), pp. 1095–1104.

[13] Panduranga S Rao et al. "A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index". In: *Transplantation* 88.2 (2009), pp. 231–236.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should I trust you?" Explaining the predictions of any classifier". In: *Proc. of the 22nd ACM SIGKDD conference*. 2016, pp. 1135–1144.

[15] Douglas E Schaubel et al. "Analytical approaches for transplant research, 2004". In: *American journal of transplantation* 5.4p2 (2005), pp. 950–957.

[16] Eric S. Weiss et al. "Creation of a Quantitative Recipient Risk Index for Mortality Prediction After Cardiac Transplantation (IMPACT)". In: *The Annals of Thoracic Surgery* 92.3 (2011), pp. 914–922.