Need for Quality Auditing for Screening Computational Methods in Clinical Data Analysis, Including Revise PRISMA Protocols for Cross-Disciplinary Literature Reviews

Julia Sidorova(✉) and Juan Jose Lozano

Centro de Investigación Biomédica en Red Enfermedades Hepáticas y Digestivas (CIBEREHD), Madrid, Spain
julia.a.sidorova@gmail.com

Abstract. Deep learning (DL) is a leading paradigm in ML, which recently has brought huge improvements in benchmarks and provided principally new functionalities. The shift towards the deep extends the horizons in seemingly every field of clinical and bioinformatics analysis. Computational platform are exposed to a great volume of new methods promising improvements. Yet, there is a trade-off between the number of man/hours and the degree to which cutting edge advances in methodology are integrated into the routine procedure. Understanding why many of the new shiny methods published in the CS literature are not suitable to be applied in clinical research and making an explicit checklist would be of practical help. For example, when it comes to survival analysis for omics and clinico-pathological variables, despite a rapidly growing number of architectures recently proposed, if one excludes image processing, the gain in efficiency and general benefits are somewhat unclear, recent reviews do not make a great emphasis on the deep paradigm either, and clinicians hardly ever use those. The consequences of these misunderstandings, which affects a number of published articles, results in the fact that the proposed methods are not attractive enough to enter applications. The example with the survival analysis motivates the need for computational platforms to work on the recommendations regarding (1) which methods should be considered as apt for a consideration to be integrated into the analysis practice for primary research articles, and (2) which literature reviews on cross-disciplinary topics are worth considering.

Keywords: survival analysis · deep learning · C-index · quality auditing · PRISMA

1 Introduction

Given the rapid advances in deep learning, for the sake of efficiency of the analysis in mission critical research, new algorithms need to be rapidly integrated into the routine clinical analysis (Sidorova, Lozano 2022a, b). There exists a known set of "regulations" regarding how machine learning must be used (Cabitza, Campagner 2021) resulting

T. Guarda et al. (Eds.): ARTIIS 2023, CCIS 1935, pp. 133–142, 2024.
https://doi.org/10.1007/978-3-031-48858-0_11

from understanding persistent errors in practical studies, as included into a checklist for the authors e.g. in the journal of Bioinformatics, but to our best knowledge there is no similar document regarding new computational methods that are to become candidates to be used in the clinical analysis. Without explicit guidelines regarding how "to separate the wheat from the chaff", computational centers need to invest time and effort into setting up the new pipelines. The protocol should save an error to the extent possible. Unfortunately, taking a publication in a premium computational journal as a suitability indication does not work, because as we will see below the objectives and criteria for success by developers are not exactly as those by practitioner. We also voice a concern regarding the PRISMA protocols (McKenzie 2021) for systematic reviews for cross disciplinary topics. For illustrative purposes, we take the topic of survival analysis based on DL.

Let us review the basics of the survival analysis to set up a technical framework for our discussion. The Cox proportional hazards (CoxPH) model (Cox 1972) for survival analysis is a well studied topic in statistics with mission critical applications in the clinical data analysis, but not so well understood by computer scientists despite the fact that a body of modern research in survival analysis now lies in deep learning. The main characteristic of the underlying mathematical problem, is that the variable of interest, Y, the time until an event (e.g. death), is attributed with an important complication of *censoring,* namely, the true event times are typically not known for all patients, because the follow up is not long enough for the event to happen, or the patient leaves the study before its termination. Due to censoring it is not correct to treat the problem as a regression. For each individual there is a true survival time T and a true censoring time C, at which the patient drops out of the study or the study ends, and

$$Y = min(T, C).$$

The status indicator is available for every observation: $\delta = 1$, if the event time is observed, and $\delta = 0$, if censoring time is observed. Furthermore, every data point has a vector of p features associated to it, also termed as attributes or covariates. The objective is to predict the true survival time T, while modeling it as a function of covariates. The (potentially stringent) proportional hazard assumption states that the hazard function has the form of

$$h(t|x_i) = h_0(t)exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p),$$

where $h_0(t)$ is a constant, called the basic hazard and no assumption is made about its functional form. The diminishing with time probability of surviving is expressed via the decreasing survival function $S(t) = Pr(T > t)$. The Kaplan-Meir is the estimator of the survival curve. For example, Fig. 1 results from learning the separation boundary between two groups of patients: those at high-risk and those with moderate risk, i.e. defining the biomarker with a sufficiently high C-index, and then plotting the $S(t)$ of the two groups. A formal test for the equality of two survival curves is the log-rank test. The CoxPH is a quite flexible modeling of the relationship between risk and covariates, robust to some violations of the initial assumptions. The conditions under which CoxPH can be correctly used are verified via statistical tests, and when unmet, other methods are recommendable, for example, see (Kleinbaum, Klein 2012) for classical extensions

or the emerging deep survival literature for DL-based alternatives (Sidorova, Lozano under review).

In trend with other fields of data science, the expectations from the DL-based survival is that it shatters benchmarks with neural networks (Higher C-index). In the context of modern hardware and flexible software frameworks, the research community revisited the idea of Faraggi-Simon (Faraggi, Simon 1995] to approximate the risk hazard *h(x)* directly with a NN, which looked promising since some potentially limiting assumptions of the CoxPH would be relaxed. The old failure to outperform CoxPH was explained with the lack of infrastructure and the under-developed theoretical apparatus. The arguments for success included:

1. Cox is linear and therefore can not learn nonlinear relations between the attributes (Huang et al. 2019; Kim et al. 2020), while there is the inherent capability of NN for efficient modeling of high-level interactions between the covariates (Katzman et al. 2018; Yang et al. 2019; Huang et al. 2019; Ching et al. 2016),
2. CoxPH relies on strong parametric assumptions that are often violated (Lee et al. 2018),
3. the desire to avoid the feature selection step (Katzman et al. 2018; Yang et al. 2019) that would lead to primitive modeling with a subsequent loss of information coded by the discarded attributes.

The primary research articles contain a spectrum of solutions ranging from a simple feed-forward network as in the first work reconsidering the idea of Faraggi and Simon (Fig. 1) to quite complex networks including a stand alone coding or classification problem, – gradually reflecting the advances in DL. When it comes to statisticcal journals, the recent quality reviews of the survival analysis do not put much emphasis on the DL-based methods, e.g. (Salerno, Li 2023), and (Lee, Lim 2019) both only briefly describe very few deep methodologies, and instead focus on the important topic of regularization and competing/semi-competing risks. Such lack of coverage can be explained with the proportional impact of the methods in real-life analysis. Despite a "Dictionary" of deep architectures was compiled with the uniform graphical representation of the main ideas and brief description of the methods (Sidorova, Lozano, under review), it comes with a series of warnings and a rather big emphasis on future work rather than on an intermediate applicability and superiority of the deep paradigm in survival, where one of the reasons why the ideas gain the grounds very slowly lie in the misconceptions regarding the success metric (typically C-index). The two statements below are insufficient as a motivation for the integration of a new method in the routine analysis:

I. "The proposed method outperforms with C-index of CoxPH with 0.05 and reaches 0.65"(Statement 1) and
II. C-index has not revealed any improvement over CoxPH but a statistically significant improvement has been detected via another metric (Statement 2).

The overly optimistic view is reflected in an *application-oriented* journal (Deepa, Gunavahi 2022) (with the correcting note submitted to the same journal (Sidorova, Lozano under review)), to which it should be added that the primary research that it cites was carried out according to the state of the art practices and published at premium venues, as well as the systematic review was designed according to PRISMA. There is a

methodological reason why the flaw crept in a cross-disciplinary review: from technical sciences to biophysics and molecular biology.

The rest of the article is organized as follows. Section 2 addresses the listed above misconceptions from the published research regarding the assessment of new algorithms for survival analysis based in DL. Section 3 discusses what is to be blamed for the misinterpretation and what can be done to save similar errors in the future. A PRISMA extension is suggested. Section 4 draws conclusions on the need for explicit set of criteria for the new methods before they can be recommended in clinical data analysis.

2 Success Metrics for Survival Analysis

The central success metric in the survival analysis is the C-index, which estimates the probability that for a random pair of individuals the predicted survival times have the same ordering as their true survival times. $C = 0.5$ is the average C-index of a random model, whereas $C = 1$ corresponds to the perfect ranking of event times, and high values of $C > 0.8$ are desired to prove the validity of a new clinical biomarker. There is an intimate relation between the area under ROC curve (AUC) and the C-index (Heagerty, Zheng 2005; Antolini et al. 2005).

2.1 Example: Misconception Regarding the Values of the Success Value in the Survival Analysis

Regarding Statement 1, unfortunately, in biomarker research typically high values of $C > 0.8$ are desired to prove the validity of a new clinical biomarker. Consider a parallel to the prediction task: one would not be willing to accept a predictive model for a binary classification with ROC of 0.6 as a sufficiently strong biomarker to be included as a test in clinical routine. Therefore, we suggest to rise *0.5* to *0.8,* for the newly proposed method to be practically valid to the a practitioner.

Although the above seems to be intuitive, as can be seen in Table 1 below, the same idea seems to be adhered in the computer science literature, where the vast majority of the proposed new architectures report rather low C-indices as long as they are greater or equal to the C-index of the CoxPH with the only exception highlighted with the bold font. The datasets SUPPORT, METABRIC, Rot&GBSD are summarized in (Kwamme et al. 2019), and the TCGA stands for the Cancer Genome Atlas (Grossman et al. 2016). The datasets are open access, which facilitates the comparison of the methods. Yet, a step is still missing to demonstrate that the proposed methods can be of practical benefit.

3. Given no improvement in C-index, with the same motivation of questionable practical benefit, an improvement detectable via a different success metric should not serve as a justification of the superiority of the method. Unless the semantics of the application implies the conditions under which the C-index can not be used.

Practitioners are left unsure with regard to the deep alternative to survival. Some articles with new clinical findings state that they are aware of these advanced deep survival methods but apologetically decline using them or describe their benefits in the Future Work. The algorithms below (DeepSurv, SALMON, and VAE-Cox) include the methods of survival analysis based on the DL that were subjectively appealing to

Table 1. C-indices for the validation of the DL-based architectures.

Study	Dataset	C-index
DeepSurv Katzman et al. (2018)	SUPPORT, METABRIC, Rot&GBSD	0.62 – 0.68
Cox-CC-Time Kvamme et al. (2019)	SUPPORT, METABRIC, Rot&GBSD	0.62 – 0.67
DeepHit Lee et al. (2018)	METABRIC	0.69
Concatenation autoencoders Tong et al. (2020)	Breast cancer data (BRCA) from TCGA, modalities: gene expression, miRNA, DNA metilation, and copy number variations	0.64
VAE-Cox (Kim et al. 2020)	10 data sets from TCGA, those with at least 50 deaths, gene expression	C = 0.65
Cox-PASNet (Hao et al. 2018)	GBM from TCGA, gene expression	C = 0.64
SurvivalNet Yosefi et al. (2017)	GBM, BRCA, KIPAN from TCGA with different set of features: 1) 17K gene expression features, and 2) the set including 3–400 clinicopathological attributes, mutations, gene- and chromosome arm-level copy number variations, and protein expression	**C > 0.8**

us: mathematical rigor, a large citation count, the treatment of the problem in the way needed, e.g. the interpretation and visualization mechanisms, etc.

1) Despite DeepSurv (Katzman et al. 2019) being a highly cited article in the literature devoted to the development in deep survival methodology (approaching 1K citations, according to Google Scholar in March 2023), less than five of them report the uses of the (parts of the) methodology in the routine clinical analysis: either directly as a method e.g., or taking a part of it with e.g. (Sahu et al. 2023).
2) In late 2022-early 2023 we have found no uses in clinical routine of SALMON (Huang et al. 2019), and yet the authors of several articles state that they are aware of this method and could have applied it, e.g. (Hu et al. 2019).
3) VAE-Cox (Kim et al. 2020) was not used in clinical studies at the time of the manuscript submission.

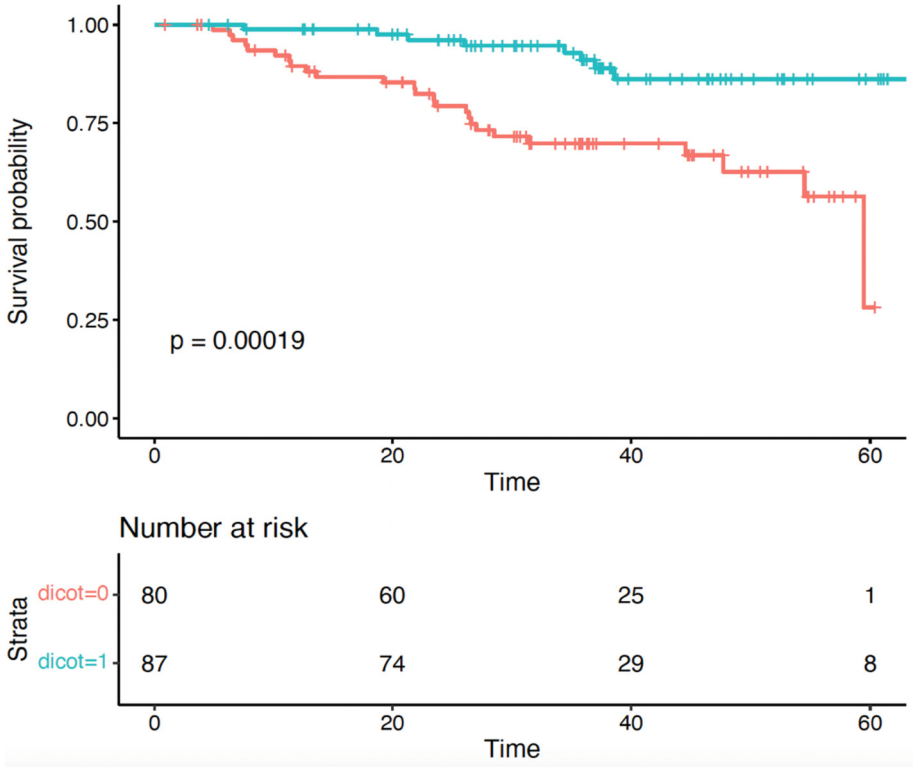

Fig. 1. The KM survival curves for the high-risk (red) and low-risk (blue) groups. Below are the counts of the survived ("at risk") patients.

3 Problem with PRISMA

The misunderstanding regarding the practical utility of the deep paradigm in the survival analysis (a warning would be suitable that these methods are hardly ever used in clinical research and that almost never an absolute high value of the success metric was reported) has crept into a non-technical journal, for which the PRISMA protocol to secure the review quality was followed. The problem is that a statistically significant improvement in the success metric, e.g. 0.05, over the baseline by CoxPH, e.g. with the resulting C-index equal to 0.65, can justify a publication in a premium CS venue, yet it is not a sufficient proof that the method is beneficial to routine clinical analysis in place or together with the state of the art methods, since to prove the discriminatory capacity of a biomarker C-index of 0.8 is desired. The systematic review in question correctly summarizes that the new methods outperform CoxPH. Let us have a look at the PRISMA flow chart (Fig. 2) to locate the place of the missing block, the aim of which will be to keep a trace of the discipline of the primary research article to be able to correctly translate the conclusions between the disciplines. We call for a need of PRISMA protocol extension along the following lines.

- *For cross-disciplinary articles, check whether the objectives behind the success metrics are the same across the disciplines of the retrieved articles, otherwise provide an explicit discussion.*

E.g. for CS research the objective is to improve the baseline as a demonstration of a suitability of the proposed algorithmic design. For clinical analysis, the objective is that the method is maximally discriminative to prove a biomarker with a typically high C-index of at least 0.8.

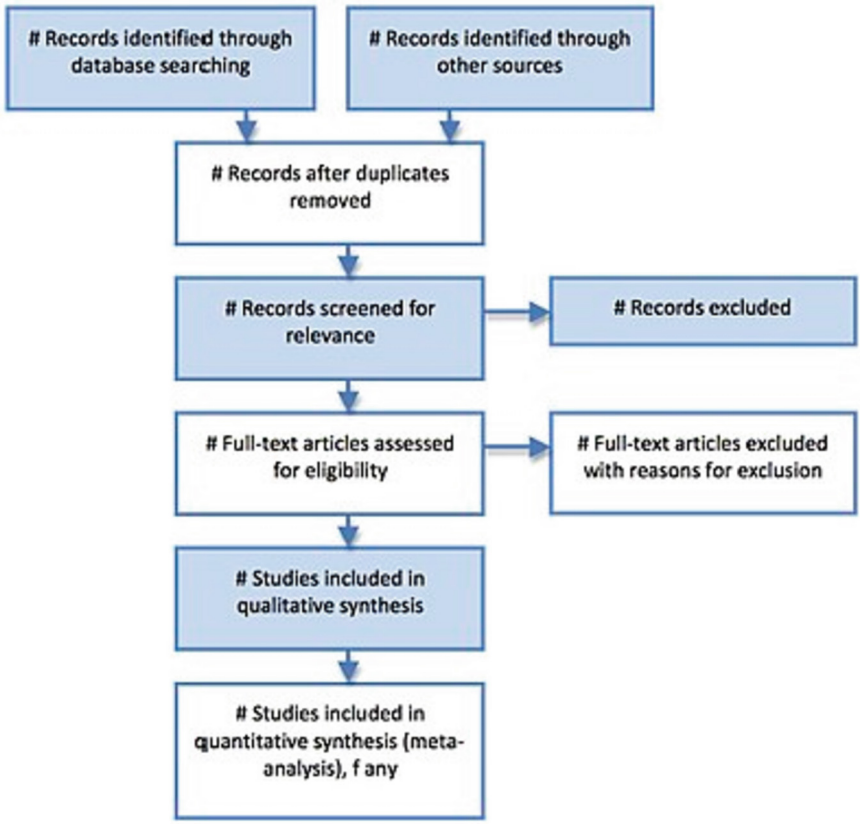

Fig. 2. The PRISMA flowchart.

Proposed Extension to PRISMA Protocol:

1. The discipline to which a primary research article belongs can be stored at the initial stage of database searching.
2. At the last block (studies included for quantitative synthesis), the articles of the same specialization should be summarized together.

3. Conclusions should be interpreted according to the research tradition within their field, and then, translated into a different discipline with an explicit discussion of the required correction/adaptation or explicitly stating no need of any change.

4 Discussion and Conclusions

The article is not aimed to promote the idea that deep survival necessarily is a dead end, moreover, in a recent article we argue for the opposite [Sidorova, Lozano, under review] yet with some reservations and wishful thinking for Future Research.

One can argue that the example falls into the "clinical utility" of the survival metric from the triple {discrimination, calibration, utility} (Hond et al. 2023). Certainly, the last metric is useful but unfortunately is largely unsed/unknown to the practitioners of survival analysis, e.g. not appearing in tutorials or literature reviews. Even if it would be. It is beneficial to assume a broader perspective and admit that a mechanistic fusion of research articles from different disciplines or mechanistic projections of the goals between disciplines is a potential source of errors in study design, e.g. see future work regarding data collection protocols.

The above example motivates the work on a systematic basis to create a set of explicit recommendations for computational methods that may be considered for routine use, – to save effort and time for the practitioners, who need to quickly filter the rapidly arriving modern algorithms that promise so much. Systematic reviews would be of much help, if they are made within the research discipline the practitioner works or if they are done with understanding of the specifics of and differences between disciplines.

If a method has no routine protocol in the main stream clinical venues. The check up list for a new method to be considered for routine application should include:

- Published at a peer reviewed technical venue with proven quality (e.g. ISI, or central CS conference such as IEEE or A/B/C-level).
- Significantly outperforms the state-of-the art method with regard to the success metric of interest as is used in the clinical literature, not any other success metric.
- The reported performance of its validation meets the standard for a publication in a clinical venue (e.g. AUC of 0.65 is not a strong argument for a new biomarker in terms of its discriminative capacity).
- If a systematic literature review based on PRISMA was used, then the cross-disciplinary differences with respect to "success" (depends on the objective of the research field) need to be the same or brought to the same scale. An explicit discussion is required on the matter.

Future Work

Unfortunately, many other subtle errors exist that make the effort invested into the application of a published and cited method results in a waste of time and effort, including the subtle differences in data collection protocols. Examples are (1) relatively easy to detect emotion expression in voice (Sidorova, Badia 2008) can be confounded with another event of interest (Sidorova et al. 2020a) such as the swing of blood glucose reflected in a vocal biomarker, (2) short-term patterns can be confounded with long-term patterns of the same disease (Sidorova et al. 2020b), and so on.

- There should be no difference between the data collection protocol in the clinical studies and the protocols adapted for the collection of data to test the algorithm, incl. The ground truth labeling. Any such minor difference must be explicitly discussed. For example, if one attempts to develop a vocal biomarker for blood glucose value, then first provide a speech sample and then read glucose value and not or vice versa. (In the case of the incorrect order, an emotion can be added to the speech sample, as the user is not entirely indifferent to one's glucose values.)

References

McKenzie, T., et al.: Statement: an updated guideline for reporting systematic reviews. BMJ **372**, 2021 (2020)

Antolini, L., et al.: A time-dependent discrimination index for survival data. Stat. Med. **24**, 3927–3944 (2005)

Cabitza, A., Campagner, A., The need to separate the wheat from the chaff in medical informatics, Introducing a comprehensive checklist for the (self) assessment of medical AI studies. Int. J. Med. Inform. **153** (2021)

Ching, T., et al, Cox-nnet: an artificial neural network method for prognosis prediction of high-thoughput omics data. PloS Comput. Biol. **14**(4) (2016)

Cox, D.R, Regression models and life-tables. J. R. Stat. Soc. Series B Methodol. **34**, 187–220 (1972)

Deepa, P., Gunavathi, C.: A systematic review on machine learning and deep learning techniques in cancer survival prediction. Progress Biophys. Molecul. Biol. **174** (2022)

Sidorova, J., Lozano, J.J.: Appendix to "A systematic review on machine learning and deep learning, under review

techniques in cancer survival prediction": Validation of Survival Methods

Faraggi, D., Simon, R.: A neural network for survival data. Stat. Med. **14**(1), 72–73 (1995)

Grossman, R.L.: et al.: Toward a shared vision for cancer genomic data. N. Engl. J. Med. **375**(12), 1109–1112 (2016)

Hao, et al.: Cox-PASNet: pathway-based sparse deep neural network for survival analysis, IEEE International Conference on Bioinformatics and Biomedicine, pp. 381–386, BIBM-2018

Heagerty, P.J., Zheng, Y.: Survival model predictive accuracy and ROC curves. Biometrics **61**, 92–105 (2005)

Hu, F., et al.: A gene signature of survival prediction for kidney renal cell carcinoma by multi-omic data analysis. Int. J. Mol. Sci. **20**(22), 5720 (2019)

Huang, et al.: SALMON: survival analysis learning with multi-omics neural networks on breast cancer. Front. Genet. (2019)

Sidorova, J., Lozano, J.J.: New and classical data processing: fruit and perils, poster. CIBER (2022)

Lee, S., Lim, H.: Review of statistical methods for survival analysis using genomic data. Genom. Inform. **17**(4), e41 (2019)

Lee Ch., et al.: DeepHit: a deep learning approach to survival analysis with competing risks. In: Proceedings of 32nd AAAI Conference on Artificial Intelligence (2018)

Kim, S., et al.: Improved survival analysis by learning shared genomic information from pan-cancer data. Bioinformatics **36** (2020)

Kleinbaum, D.G., Klein, M.: Survival analysis, a self-learning text, 3rd edn. Springer, Statistics for Biology and Health (2012)

Kvamme, H., et al.: Time-to-event prediction with neural network and Cox regression. J. Mach. Learn. Res. **20** (2019)
Katzman, J., et al.: DeepSurv: personalised treatment recommender system using a Cox proportinal hazards deep neural network. BMC Med. Res. Methodol. **18**, 24 (2018)
Sahu, et al.: Discovery of targets for immune-metabolic antitumor drugs identifies estrogen-related receptor alpha. Cancer Discov. **13**(3), 672–701 (2023)
Salerno, S., Li, Y.: High-dimensional survival analysis: methods and applications. Ann. Rev. Statist. Appl. **10**, 25–49 (2023)
Sidorova, J., Lozano, J.J.: New and classical data processing: fruit and perils, jornadas Cientificas, CIBER (2022)
Sidorova, J., Lozano, J.J.: A Survey of Survival Analysis with Deep Learning: Models, Applications and Challenges. under review
[SEER, http] https://seer.cancer.gov/causespecific/
Tong, L., et al.: Deep learning based feature-level integration pf multi-omics data for breast cancer patients survival analysis. BCM Med. Informa. Dec. Mak. **20**, 225 (2020)
Yang, et al.: Identifying risk stratification associated with a cancer for overall survival by deep-learning based CoxPH. IEEE Access (7) (2019)
Yousefi, S., et al.: Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Scientific Reports (2017)
Sidorova, J., et al.: Blood glucose estimation from voice: first review of successes and challenges. J. Voice (2020a)
Sidorova, J., et al.: Impact of diabetes mellitis on voice: a methodological commentary (2020b)
Hond, A., et al.: Perspectives of validation of clinical predictive algorithms, digital Medicine (2023)
Sidorova, J., Badia, T.: ESEDA: tool for enhanced emotion recognition and detection, Procs. AXMEDIS (2008)