Application of machine learning techniques in the optimization of produced water treatment: A literature review

Diogo Souza Neiva Cardoso[1,*], Kleberson Ricardo de Oliveira Pereira[1], George Simonelli[1], Luiz Carlos Lobato dos Santos[1]

[1]*Oil, Gas, and Biofuels Research Group, Postgraduate Program of Chemical Engineering, Federal University of Bahia. Salvador, Bahia, Brazil.*

**E-mail: dsncardoso.eq@gmail.com*

Abstract: The increasing complexity and variability of operational and physicochemical conditions in produced water treatment within the oil and gas industry require advanced computational solutions capable of managing multivariate data and modeling nonlinear behaviors. In this context, machine learning (ML) techniques have emerged as powerful tools for predicting critical parameters, optimizing treatment processes, and supporting decision-making. This study presents a critical review of publications from 2015 to 2025, focusing on the application of ML in the optimization of produced water treatment. The literature search was conducted in the Scopus database using specific descriptors and Boolean operators, and resulted in 81 selected studies. The number of publications has grown significantly since 2020, with the United States, China, Canada, and Brazil leading the research effort. The predominant techniques include artificial neural networks (ANNs), support vector machines (SVM), and ensemble models such as Random Forest and XGBoost, which showed high predictive accuracy ($R^2 > 0.90$ in most studies). Key input variables included operational parameters, ionic composition, and historical production data. Reported outcomes include improved removal efficiency, scale control, and operational cost reduction of up to 30%. Despite advances, limitations remain regarding overfitting risks, lack of long-term validation, and model interpretability. Overall, the findings suggest that ML techniques represent a promising approach to enhance produced water treatment, enabling more efficient, adaptable, and economically viable industrial applications.

Keywords: Process optimization, Oil industry, Artificial neural networks, Random Forest, XGBoost.

1. Introduction

Produced water (PW) stands out as one of the major environmental and operational challenges faced by the oil industry, due to its large volumes and the variable composition of contaminants such as oils, greases, salts, suspended solids, and dissolved organic compounds [1,2,3]. This heterogeneous composition complicates the stable operation of treatment systems and undermines the predictability of key performance indicators, such as oil removal efficiency, chemical oxygen demand (COD), and turbidity [2,3].

Traditional modeling methods, based on empirical correlations, linear regressions, or classical statistical models, show significant limitations in handling multivariate processes and nonlinear behaviors, which are characteristic of real PW treatment systems. Although useful for understanding general trends, these approaches struggle to capture the complex interactions among variables that frequently occur in industrial operations subject to fluctuations in flow rate, composition, and operational conditions over time [3,4,5].

In this context, machine learning techniques have emerged as promising tools to overcome these limitations by enabling the modeling of complex nonlinear relationships, identifying the most influential variables, and optimizing operational parameters. Models such as artificial neural networks (ANNs), support vector machines (SVMs), decision trees, and ensemble methods, including Random Forest and XGBoost, have been applied to predict the removal efficiency of

specific contaminants, adjust chemical dosing, and reduce energy consumption [3,6]. By offering higher predictive accuracy and adaptability to different operational scenarios, these techniques directly contribute to improving treatment system efficiency, ensuring compliance with environmental standards, and optimizing costs within the industrial context of produced water [1,5]. The adoption of such approaches has intensified in recent years, reflecting a global trend toward integrating computational intelligence into complex environmental processes.

Accordingly, this review study aims to critically assess the application of machine learning techniques in optimizing produced water treatment, highlighting the types of models employed, the most common input and output variables, the reported performance metrics, and the main limitations identified in the literature.

2. Methodology

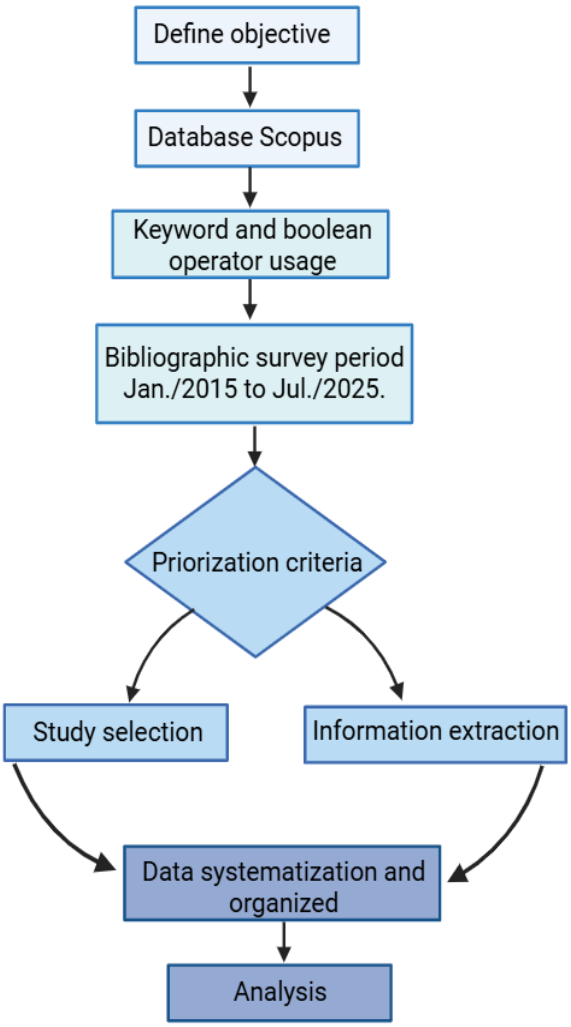
This literature review was conducted in a critical and interpretative manner, aiming to analyze the use of machine learning techniques in the modeling and optimization of produced water (PW) treatment processes. The search was carried out in the Scopus database using descriptors combined with Boolean operators: ("produced water" OR "oilfield produced water" OR "petroleum wastewater") AND ("machine learning" OR "artificial neural networks" OR "ANN" OR "support vector machine" OR "decision trees" OR "random forest" OR "XGBoost" OR "deep learning") AND ("treatment" OR "removal" OR "prediction" OR "optimization"). The bibliographic survey covered the period from January 2015 to July 2025, prioritizing articles published in indexed journals that presented practical applications or quantitative results related to PW treatment.

The choice of the Scopus database is justified by its status as one of the leading references for researchers worldwide, as well as its extensive collection of relevant studies on the topic. Scopus was also used to generate the bibliometric overview. From this database, indicators such as the annual volume of publications, geographic distribution by country/territory, and types of documents published during the selected period were collected and analyzed.

Studies applying various machine learning models, including artificial neural networks (ANNs), support vector machines (SVMs), and ensemble methods (Random Forest, XGBoost), were selected, focusing on predicting contaminant removal efficiency, optimizing operational parameters, or reducing costs and energy consumption. For each article, information was extracted regarding the type of model, input and output variables, performance metrics, such as the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute percentage error (MAPE), as well as the main findings reported by the authors. The collected data were systematized in a spreadsheet and organized into a comparative table, enabling a descriptive and exploratory analysis of the most used algorithms, studied

variables, and obtained results. Figure 1 illustrates the methodological procedure employed in this study.

Figure 1. Methodological procedure used to conduct this study.

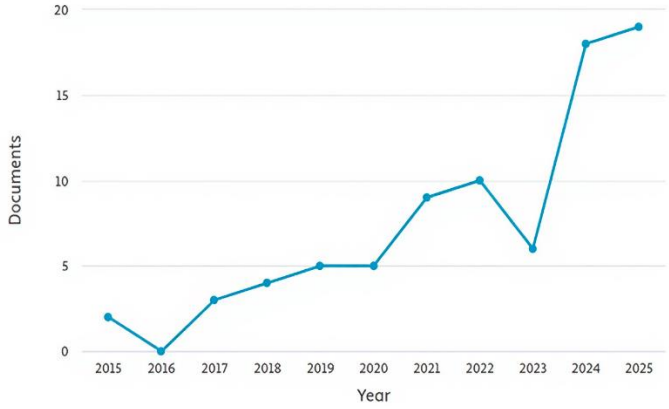


3. Results and Discussion

3.1 Evolution and Profile of Publications

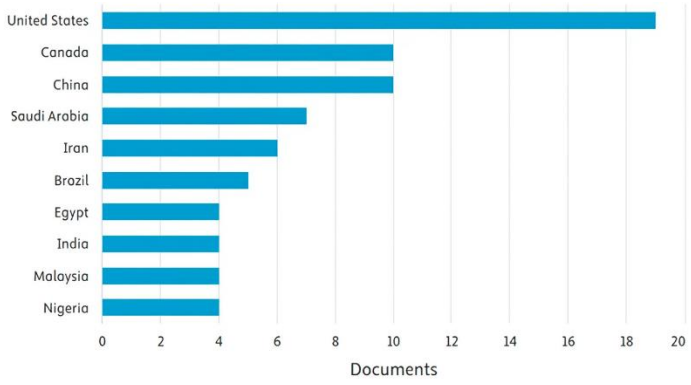
The evolution of publications related to the application of machine learning in produced water treatment is presented in Figure 2. Data extracted from the Scopus database totaled 81 publications.

Figure 2. Publications per year.



A gradual increase is observed until 2022, followed by a more pronounced rise between 2023 and 2024, with the upward trend continuing into 2025, considering that the data were collected up to July of this year. This recent growth may be attributed to greater accessibility of ML tools, advancements in algorithms, and the growing interest in process automation and sustainability in treatment systems. Figure 3 shows the geographic distribution of the published studies.

Figure 3. Publications by country/territory.

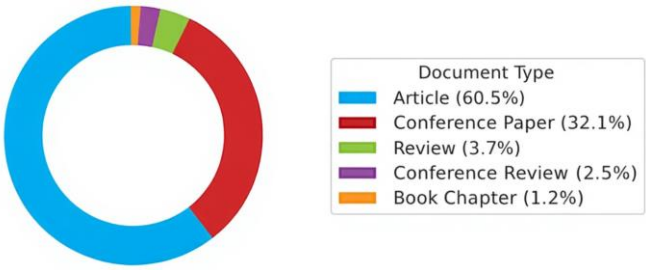


The United States leads in the number of studies, reflecting its prominent role in adopting emerging technologies within the oil industry and

environmental management. Following are Canada and China, as well as Middle Eastern countries (Saudi Arabia and Iran), where high volumes of produced water are generated due to oil exploration. Brazil holds an intermediate position with five publications, indicating growing interest in optimization solutions for produced water treatment. Other countries such as Egypt, India, Malaysia, and Nigeria show targeted initiatives that underscore the global reach of the topic and its potential application across diverse production contexts.

Regarding document types (Figure 4), journal articles predominate (60.5%), followed by conference papers (32.1%). This distribution reflects the consolidation of the topic within the scientific literature and the growing interest of the academic community in both applied studies and exploratory approaches.

Figure 4. Types of published documents.



Additionally, reviews, book chapters, and conference reviews appear in smaller proportions but play a strategic role in the systematization of knowledge, particularly in establishing conceptual frameworks, identifying research gaps, and proposing future research directions. The combination of these different types of publications reflects an expanding field, marked by solid advances and, at the same time, openness to further investigation.

3.2 Machine learning techniques in the optimization of produced water treatment

Recently, machine learning techniques have gained prominence in the oil and energy sectors, particularly due to their ability to handle complex data and support the optimization of industrial processes [4,11]. In produced water (PW) treatment, this progress is reflected in a growing number of studies applying predictive models to enhance operational efficiency, reduce costs, and comply with increasingly stringent environmental regulations [6,13].

The choice of the most appropriate technique is closely linked to the nature of the problem, the availability and quality of the data, and the specific objectives of each application. This methodological diversity reflects the scientific community's effort to balance accuracy, interpretability, and industrial applicability [6,10,13]. When properly trained and validated, such models exhibit strong potential to anticipate critical behaviors and support decision-making, even under operational fluctuations and uncertainties [10,11].

Table 1 presents these studies, detailing the types of machine learning models employed, input and output variables, reported performance, and their respective reference.

Table 1: Studies employing machine learning in the optimization of produced water treatment.

Machine learning	Main Objective	Input Variables	Output Variables	Performance (R^2 / MAPE)	Ref.
MLR, RF, XGBoost	Predict scale formation	Concentrations of Ca^{2+}, Mg^{2+}, Ba^{2+}, Sr^{2+}, SO_4^{2-}, HCO_3^-, CO_3^{2-}, Cl^-, temperature, and pH	Scaling index	**XGBoost**: R^2=0.95	[7]
SVM, RF, XGBoost, etc. (8 models)	Predict desalination efficiency	Salt concentration, TDS, pH, temperature, pressure	Removal efficiency (%)	**SVM**: R^2=0.98	[8]
ANN	Predict oil growth rate	Daily processing capacity, oil content, polymer concentration, suspended solids	Oil growth rate	**ANN**: R^2≈0.83–0.91	[9]
ANN, XGBoost	Optimize membrane treatment	Temperature, flow rate, shear rate, pressure	Transmembrane pressure (TMP) and total resistance	**ANN**: R^2=0.95 **XGBoost**: R^2=0.97	[10]
Linear regression, KNN, ANN	Improve efficiency and reliability of filtration systems	Droplet radius, time, shear rate	Lift force, TMP, resistance	**Força de elevação**: R^2>0.99 **TMP (KNN)**: R^2=0.90	[11]
RFR, ARIMA	Predict produced water volume and quality	Well age, production rate, location, geological data	Produced water volume	**RFR**: R^2>0.80; **ARIMA**: MAPE≈5.5%	[12]
ANN	Optimize treatment costs	Pressure, salt concentration, temperature, production rate	Cost, removal efficiency, coagulant dosage, etc.	**Qualitative analysis**	[13]
ANN, RSM	Optimize turbidity removal	Dosage, temperature, time	Removal efficiency (%)	**ANN**: R^2=0.99 **RSM**: R^2=0.99	[14]
ANN	Model biological polishing unit	Temperature, pH, COD, suspended solids	Effluent COD	**ANN**: R^2=0.84	[15]
ANN	Predict ion rejection in membranes	pH, pressure, flow rate	Ion rejection (%): Cl^-, Na^+, Mg^{2+}, Ca^{2+}	**ANN**: R^2>0.95	[16]

ANN (*Artificial Neural Network*)
MLP (*Multilayer Perceptron*)
SVM (*Support Vector Machine*)
RF (*Random Forest*)
XGBoost (*Extreme Gradient Boosting*)
RFR (*Random Forest Regression*)
ARIMA (*AutoRegressive Integrated Moving Average*)
RSM (*Response Surface Methodology*)
RMSE (*Root Mean Square Error*)
MAPE (Mean Absolute Percentage Error)/
MLR (Multiple Linear Regression

The findings of this study highlight diverse applications of machine learning (ML) techniques in produced water treatment, encompassing artificial neural networks (ANNs),

support vector machines (SVMs), ensemble algorithms (Random Forest, XGBoost), and classical statistical models such as ARIMA. The most common input variables included operational parameters (temperature, flow rate, pressure), ionic composition (Ca^{2+}, Mg^{2+}, Ba^{2+}, Sr^{2+}, Cl^{-}, SO_4^{2-}, etc.), physicochemical characteristics (salinity, oil content, suspended solids, COD), and, in some cases, geological data or historical production records. Some datasets were obtained from laboratory experiments or pilot plants, while other studies, such as Wang et al. [9] and Jiang et al. [12], utilized field time series data. Output variables primarily focused on predicting contaminant removal efficiency, ion rejection, oil growth rate, operational costs, and produced water volume. The ML models demonstrated robust performance, frequently achieving R^2 values above 0.90 with low error metrics (RMSE, MAPE).

Among the highlighted examples, different machine learning techniques were applied to specific and complex challenges in produced water treatment, always involving multiple correlated variables and nonlinear behavior. Tayyebi et al. [7] modeled scale formation based on ion concentrations (Ca^{2+}, Mg^{2+}, Ba^{2+}, Sr^{2+}, SO_4^{2-}, HCO_3^{-}, CO_3^{2-}, Cl^{-}), temperature, and pH, achieving an R^2 of 0.95 with XGBoost, which stood out for its ability to capture complex interactions among geochemical variables; however, the authors noted the need for constant calibration due to variations in water composition over time. Nallakukkala et al. [8] reached an R^2 of 0.98 using SVM to predict desalination efficiency via gas hydrates, highlighting the technique's capability to model nonlinear relationships between operational variables such as TDS, pH, temperature, and pressure; nevertheless, they acknowledged limitations related to the scarcity of experimental data at an industrial scale. Additionally, Wang et al. [9], predicting oil growth rate in settling tanks using ANN, reported R^2 values ranging from 0.83 to 0.91 but emphasized dependence on fragmented historical data and the risk of overfitting.

Saddiqi et al. [10] demonstrated that ANNs and XGBoost achieved high R^2 values (0.95–0.97) when modeling transmembrane pressure and total resistance in membrane processes, while Saddiqi et al. [11] extended the modeling to hydrodynamic variables such as lift force, droplet radius, and shear rate, achieving $R^2 > 0.99$. These results reinforced the critical role of hydrodynamic variables in reducing fouling. Ezemagu et al. [14] showed that ANNs outperformed the response surface methodology (RSM) in predicting turbidity removal efficiency ($R^2 = 0.99$ versus 0.97), demonstrating a greater ability to model nonlinear relationships.

Nair et al. [16] combined ANNs with physicochemical models (Spiegler-Kedem and SHP) to predict ion rejection by nanofiltration, achieving R^2 values above 0.95; however, they cautioned about the need for long-term testing to assess fouling. Meanwhile, Aisyah et al. [15] applied ANN to model a biological polishing unit, obtaining an R^2 of 0.84. Although

satisfactory, the study highlighted the model's sensitivity to variations in operational conditions and the difficulty of representing complex biological processes using traditional mathematical models.

Jiang et al. [12] explored different techniques: Random Forest Regression (RFR), an ensemble tree-based method, achieved an R^2 greater than 0.80, while the classical statistical model ARIMA reached a mean absolute percentage error (MAPE) of approximately 5.5% in forecasting time series of produced water volume and quality. This combination illustrates how different methods can be complementary, depending on the nature of the data. On the other hand, Taloba [13] presented a qualitative analysis indicating that ANNs can optimize energy consumption and reduce operational costs (up to 25% less energy and 30% lower costs). Although these gains are promising, the author emphasizes that industrial application requires broader experimental validation.

Thus, it is observed that multilayer perceptron (MLP) architectures of ANNs, combined with ensemble algorithms such as XGBoost and Random Forest Regression (RFR), predominate due to their ability to capture complex nonlinear relationships and handle multiple variables. Nonetheless, challenges remain: the risk of overfitting with limited datasets [2,11,17], the need for field validation, and improved model interpretability, which are fundamental aspects to enable large-scale applications.

Collectively, these studies demonstrate that by considering different types of variables and employing techniques suited to each case, the use of machine learning significantly enhances the ability to anticipate critical process behaviors, overcome the limitations of traditional models, and support operational decision-making in complex industrial scenarios. Therefore, the results suggest that applying machine learning to optimize produced water treatment offers gains in predictive accuracy, operational efficiency, and cost reduction provided it is supported by comprehensive datasets, long-term testing, and appropriate validation strategies.

4. Conclusion

This literature review addresses recent advances in the use of machine learning techniques to enhance produced water treatment processes. The following conclusions are drawn:

- There has been a marked increase in publications in recent years, indicating growing global interest in computational solutions applied to produced water management in the oil industry;
- Machine learning techniques overcome limitations of traditional statistical models by effectively modeling nonlinear and multivariate relationships typical of treatment processes;
- Artificial neural networks (ANNs), combined with ensemble algorithms such as XGBoost and Random Forest Regression (RFR), demonstrate

higher predictive accuracy, with R^2 values exceeding 0.90 in most analyzed cases;

- The incorporation of operational, physicochemical, and historical production data enhances the ability to anticipate critical process behaviors and support decision-making in complex industrial scenarios;

- The reviewed studies demonstrate applications such as prediction of scale formation, estimation of contaminant removal efficiency, analysis of operational costs, and prediction of ion rejection in membranes, highlighting machine learning's role as a powerful tool for monitoring and optimizing produced water treatment.

Acknowledgments

The authors gratefully acknowledge the financial support from the Human Resources Program of the National Agency of Petroleum, Natural Gas and Biofuels (PRH/ANP – PRH36/UFBA), funded through investments by qualified oil companies under the R&D Clause of ANP Resolution No. 50/2015, as well as the National Council for Scientific and Technological Development (CNPq). In addition, this study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] Abdelhamid C, Latrach A, Rabiei M, Venugopal K. Produced Water Treatment Technologies: a review. *Energies*. 2025; 18(1):63.

[2] Al-Ajmi F, Al-Marri M, Almomani F, AlNouss A. A Comprehensive Review of Advanced Treatment Technologies for the Enhanced Reuse of Produced Water. *Water*. 2024;16(22):3306.

[3] Amakiri KT, Canon AR, Molinari M, Angelis-Dimakis A. Review of oilfield produced water treatment technologies. *Chemosphere*. 2022;298:134064.

[4] Alardhi SM, Jabbar NM, Breig SJM, Hadi AA, Salman AD, Saedi LMA; Khadium MK, Showeel HA, Malak HM, Mohammed MM. Artificial neural network and response surface methodology for modeling oil content in produced water from an Iraqi oil field. *Water Practice & Technology*.2024;19(8):3330-3349.

[5] Chaal RE, Aboutafail MO. A comparative study of back-propagation algorithms: levenberg-marquart and bfgs for the formation of multilayer neural networks for estimation of fluoride. *Commun. Math. Biol. Neurosci*. 2022;1-17.

[6] Ibrahim M, Haider A, Lim JW, Mainali B, Aslam M, Kumar M; Shahid, MK. Artificial neural network modeling for the prediction, estimation, and treatment of diverse wastewaters: a comprehensive review and future perspective. *Chemosphere*. 2024;362:142860.

[7] Tayyebi A, Alshami A, Tayyebi E; Owoade A, Talukder M, Ismail N, Rabiei Z, Yu X, Tikeri G. Machine learning-based prediction of scale formation in produced water as a tool for environmental monitoring. *Results In Engineering.* 2025;26:105223.

[8] Nallakukkala S, Tackie-Otoo BN, Aliyu R, Lal B, Nallakukkala JRD, Devi G. Application of machine learning algorithms to predict removal efficiency in treating produced water via gas hydrate-based desalination. *Desalination.* 2025;612:118961.

[9] Wang Z, Wang C, Meng L, Qi X, Hong J. Prediction of sump oil growth rate towards gravitational settling of produced water in oilfield based on machine learning. *Desalination and Water Treatment*. 2024;317:100189.

[10] Saddiqi HA, Javed Z, Ali QM, Ullah A. Optimization and predictive modeling of membrane based produced water treatment using machine learning models. *Chemical Engineering Research and Design.* 2024;207:65-76.

[11] Saddiqi HA, Javed Z, Ali QM, Ullah A, Ahmad I. Modelling and predicting lift force and trans-membrane pressure using linear, KNN, ANN and response surface models during the separation of oil drops from produced water. *Journal of Water Process Engineering.* 2024; 66:106014.

[12] Jiang W, Pokharel B, Lin L, Cao H, Carroll KC, Zhang Y, Galdeano C, Musale DA, Ghurye GL, Xu P. Analysis and prediction of produced water quantity and quality in the Permian Basin using machine learning techniques. *Science of The Total Environment*. 2021;801:149693.

[13] Taloba AI. An Artificial Neural Network Mechanism for Optimizing the Water Treatment Process and Desalination Process. *Alexandria Engineering Journal.* 2022;61(12):9287-9295.

[14] Ezemagu IG, Ejimofor MI, Menkiti MC, Nwobi-Okoye CC. Modeling and optimization of turbidity removal from produced water using response surface methodology and artificial neural network. *South African Journal of Chemical Engineering*. 2021;35:78-88.

[15] Aisyah PY, Soehartanto T, Finazis RF, Afif K, Lokeswara R, Umamah F. Process Dynamics Modeling on Polishing Unit of Artificial Neural Network-Based Produced Water Treatment System. *IEEE International Conference on Advanced Mechatronics, Intelligent Manufacture And Industrial Automation (Icamimia)*. 2021. p. 23-27.

[16] Nair RR, Protasova E, Strand S, Bilstad T. Effect of pH on produced water treatment using nanofiltration membranes: artificial neural network for performance assessment and steric hindrance pore model for flux variation evaluation. *Desalination and Water Treatment.* 2019;146:120-130.

[17] Heydari B, Sharghi, EA, Rafiee S, Mohtasebi SS. Use of artificial neural network and adaptive neuro-fuzzy inference system for prediction of biogas production from spearmint essential oil wastewater treatment in up-flow anaerobic sludge blanket reactor. *Fuel.* 2021;306:121734, 2021.