Research Article

The 3Ranker: An AI-based Algorithm for Finding Non-animal Alternative Methods

Alternatives to Laboratory Animals
2023, Vol. 0(0) 1–11

DOI: 10.1177/02611929231210777
journals.sagepub.com/home/atl
Sage

Niels van Beuningen[1], Sinne Alkema[1], Nils Hijlkema[1], Brun Ulfhake[2], Rafael Frias[3], Merel Ritskes-Hoitinga[4,5] and Wynand Alkema[1,6]

Abstract

The search for existing non-animal alternative methods for use in experiments is currently challenging because of the lack of both comprehensive structured databases and balanced keyword-based search strategies to mine unstructured textual databases. In this paper we describe 3Ranker, which is a fast, keyword-independent algorithm for finding non-animal alternative methods for use in biomedical research. The 3Ranker algorithm was created by using a machine learning approach, consisting of a Random Forest model built on a dataset of 35 million abstracts and constructed with weak supervision, followed by iterative model improvement with expert curated data. We found a satisfactory trade-off between sensitivity and specificity, with Area Under the Curve (AUC) values ranging from 0.85–0.95. Trials showed that the AI-based classifier was able to identify articles that describe potential alternatives to animal use, among the thousands of articles returned by generic PubMed queries on dermatitis and Parkinson's disease. Application of the classification models on time series data showed the earlier implementation and acceptance of Three Rs principles in the area of cosmetics and skin research, as compared to the area of neurodegenerative disease research. The 3Ranker algorithm is freely available at www.open3r.org; the future goal is to expand this framework to cover multiple research domains and to enable its broad use by researchers, policymakers, funders and ethical review boards, in order to promote the replacement of animal use in research wherever possible.

Keywords

3Rs, artificial intelligence, predictive models, replacement, text classification, text mining, Three Rs

Introduction

Replacement, *refinement* and *reduction* (the Three Rs) are principles that play a crucial role in the ethical treatment of animals in scientific research. These principles aim to minimise animal suffering, reduce the numbers of animals used, improve animal welfare, and promote the responsible use of animals in experiments. The Three Rs encourage researchers to refine any experimental procedures that are carried out, in order to reduce pain and distress, and to explore alternative methods that can replace the use of animals in their experiments. However, the search for such existing non-animal alternative methods is challenging for several reasons.

Firstly, although more than 100 databases and Three Rs websites are available that feature information on the Three Rs principles, their content, annotation and search strategies vary enormously, preventing an effective cross-database search. In addition, direct links to the underlying scientific literature are not always present.[1,2] Moreover, these databases are manually updated, leading to a considerable delay between the publication of possible Three Rs strategies in the primary literature and the inclusion of these papers in the relevant databases.

Secondly, even in structured databases like PubMed/MEDLINE and Embase, identifying Three Rs-related

[1]TenWise BV, Leiden, The Netherlands
[2]Department of Laboratory Medicine, Karolinska Institute, Solna, Sweden
[3]Department of Comparative Medicine, Karolinska Institute, Solna, Sweden
[4]Department Population Health Sciences — IRAS Toxicology, Utrecht University, Utrecht, The Netherlands
[5]Department Clinical Medicine, Aarhus University, Denmark
[6]Institute for Life Science and Technology, Centre for Biobased Economy, Hanze University of Applied Sciences, Groningen, The Netherlands

Corresponding author:
Wynand Alkema, Institute for Life Science and Technology, Centre for Biobased Economy, Hanze University of Applied Sciences, Zernikeplein 11 Groningen 9700 RM, The Netherlands.
Email: w.b.l.alkema@pl.hanze.nl

papers is difficult, because these methods are often not explicitly described as 'animal alternatives' in the scientific abstracts. This hampers the creation of general unbiased keyword-based search strategies with an acceptable trade-off between specificity and sensitivity.[3]

The identification of animal replacement alternatives can also be accomplished by performing systematic reviews consisting of an exhaustive literature search of databases, followed by manual expert curation — indeed, this approach can yield suggestions for non-animal alternatives, and can lead to the successful implementation of the Three Rs principles.[4] However, given that systematic reviews are quite time-consuming and hard to keep up-to-date, they are not yet accepted as the most efficient way forward for the fast and effective identification of non-animal alternatives.

These challenges and difficulties have been confirmed by researchers, who have indicated in various questionnaires that they think it highly likely that Three Rs alternatives for their proposed research exist, but that they have not been able to find them.[1]

In recent years, a large number of text-mining software applications have been developed to assist researchers in extracting relevant publications from large amounts of search results. Text-mining software applications are aimed at the fast and unbiased retrieval of papers, classification of papers according to topic, and information extraction from the abstracts and full text of these papers. Examples include: the use of machine learning for detecting teratology papers;[5] the identification and classification of clinical texts for classifying smoking status and hip fractures;[6] and information extracted from abstracts that describe randomised trials for stroke[7] and other Randomised Controlled Trials (RCTs),[8] that was then used in systematic and scoping reviews.[9] The models that were developed in these papers were based on expert curated reference sets, which were subsequently converted into vectorised representations, and coupled to a machine learning model, such as Support Vector Machines, Convolutional Neural Networks or Random Forest. These models were subsequently evaluated by measuring the sensitivity (recall) and specificity for a number of threshold scores of the models.

Building further on these developments, in the current paper we describe the development of a Random Forest (RF) classifier for the identification of Three Rs-related papers. Recently, we reported on the use of AI to analyse Three Rs-related papers, and showed the usefulness of word embeddings to extract topics and their semantic relationships from the Three Rs literature.[10] Here, we extend this work and describe the 3Ranker — a fast, keyword-independent Three Rs search algorithm based on a RF classifier that can be used to retrieve suggestions for non-animal alternatives based on a user-specified search string. We applied this method to a large set of papers related to dermatitis and Parkinson's disease. Dermatitis was chosen in view of the EU ban on animal testing for cosmetics, which has led to a number of successful non-animal replacement skin sensitisation and irritation tests being developed, validated and even legally accepted.[11–13] Parkinson's disease was chosen as an example of a neurodegenerative systemic disease, which are believed to be generally less amenable to the application of non-animal alternative approaches.[14] Using 3Ranker, we efficiently identified papers featuring potential Three Rs alternatives among the thousands of abstracts that were retrieved for these topics. The algorithm and a user-friendly web interface for searching are available via www.open3r.org.

Materials and methods

Medline abstracts

All methods described in the paper were developed by using MEDLINE-listed abstracts as the text source. The abstracts were downloaded from the MEDLINE baseline release of 2023 (released on 8 December 2022), and after that on a daily basis from the incremental updates supplied by MEDLINE, until March 2023. All downloads were done from the MEDLINE FTP site (https://www.nlm.nih.gov/databases/download/pubmed_medline.html) in XML format, and parsed to tabular format for further processing. All abstracts were subsequently indexed with the Lucene software (https://lucene.apache.org/). All text manipulations on abstracts, such as transformation to lower case, removing stop words and punctuation, were performed with R (www.r-project.org) by using the libraries tm[15] and tidyverse.[16]

Predictive modelling

The parsed XML files were used as input for predictive modelling. To this end, all words starting with a single uppercase letter were converted to lowercase. Words with all uppercase letters (often abbreviations or acronyms) were not modified. All punctuation characters were replaced with a single space. Next, all words in each abstract were counted, resulting in a 'Bag-of-Words' (BOW) — i.e. a large matrix with rows being the abstracts and the columns representing the words and their count in each abstract. This matrix was used as input for creating RF models. Creation of the RF classifier was carried out with the 'randomForest' package in R, by using the default settings.[17] The RF classifier was trained on 70% of the positive (Three Rs) and negative (not Three Rs) dataset with the default settings for the classifier. The remaining 30% of the data was kept isolated and used for validation of the model. This procedure was repeated 20 times, each time starting with a different split of 70–30% of the total dataset. The performance

of the models was evaluated with Receiver Operator Characteristics (ROC) curves by using the pROC package in R.[18]

Expert curation

A team of six (for the SKIN model) and eight (for the BRAIN model) reviewers were presented with 1000 abstracts in total, and were asked to score each abstract's relevance to the Three Rs on a three-point scale, namely: high (2 points); medium (1 point); and low (0 points). Abstracts were presented in random order to the reviewers and each reviewer scored between 200 and 300 abstracts, meaning that not all abstracts were seen by every reviewer. To keep the process for generating the training and test sets as agile and extensible as possible, we did not attempt to resolve scoring conflicts between reviewers for specific abstracts. Instead, for each abstract, an average score was calculated by dividing the sum of the reviewer scores by the number of reviewers. Abstracts with an average score > 1.5 were regarded as positive, and abstracts with a score of 0 were regarded as negative abstracts; abstracts with a score between 0 and 1.5 were not included in model building.

Availability

The interactive web interface for the application of the final models was built in R by using the 'shiny' package (https://rstudio.github.io/shiny/index.html) and is available at www.open3r.org. The RF models and R scripts with examples for the usage of these models are available upon request from the corresponding author.

Results

Creation of the de novo model

To create predictive models for the recognition of papers featuring Three Rs alternatives, a starting model (the '*de novo* model') was firstly constructed with a dataset subjected to weak supervision. This set of abstracts was enriched for Three Rs papers, irrespective of the research domain, but not yet curated by human experts. The initial set of abstracts was compiled as follows:

1. All abstracts with the keywords *3R principle*, *3R approaches*, *3R awareness*, *3R compliant*, *3R concept*, *3R framework*, *3R method**, *3R policy*, *animal alternative* and *non-animal alternative* were collected from MEDLINE.
2. All abstracts from the journal *Alternatives to Laboratory Animals (ATLA)* were collected.
3. All abstracts with the MeSH terms *Animal Testing Alternatives* and *Animal Use Alternatives* were collected from MEDLINE.

The abstracts obtained after following the above steps yielded a non-redundant abstract collection of, in total, 4701 abstracts — in this paper, they are referred to as the *de novo* abstract set. This set was extended with a set of 5000 random abstracts, which served as a negative set when training the model. It should be noted that the goal was not to collect an exhaustive set of all abstracts related to the Three Rs, but rather to create a starting set that can be assumed to be enriched for Three Rs-related papers. This *de novo* set was subsequently used to generate the first predictive model (*de novo* predictive model) by training a RF model on the words contained in the abstracts of the *de novo* set.

The performance of the model was evaluated by the specific keywords that were identified as most predictive. These predictive features were obtained by summing the importance score assigned by the RF model in the 20 runs that were used to create the final model (Figure 1(a)) and the overall predictive performance, which is expressed by the Area Under the Curve (AUC) in ROC plots (Figure 1(b)). The AUC indicates how well the model separates the negative from the positive abstracts, where a value of 0.5 represents a model that is no better than chance, and an AUC value of 1 represents a model with 100% sensitivity and specificity at each threshold value.

Among the most predictive features (i.e. words that discriminate Three Rs papers from non-Three Rs papers) are keywords such as animal, alternative, predictive, toxicology — these are indeed keywords that may be expected to be specific, especially in combination, for papers that describe Three Rs alternatives. In addition, the models had high AUC scores, ranging from 0.8 to 0.95, indicating satisfactory predictive power. Taken together, these results show that with a non-expert-curated abstract set and standard RF classifiers, a satisfactory starting model for recognising papers featuring potential Three Rs alternatives was obtained.

Application of the de novo model

To evaluate the applicability of the classification model, the *de novo* model was used to score two sets of abstracts (Table 1). One abstract set was obtained by searching PubMed using the term 'dermatitis'; the second set was obtained by searching PubMed using the term 'Parkinson's disease'. In both cases, the search was limited to the 20,000 most recent abstracts. Each abstract was evaluated by the *de novo* model, resulting in a probability score for the abstract featuring a Three Rs alternative of between 0 (low probability) and 1 (high probability).

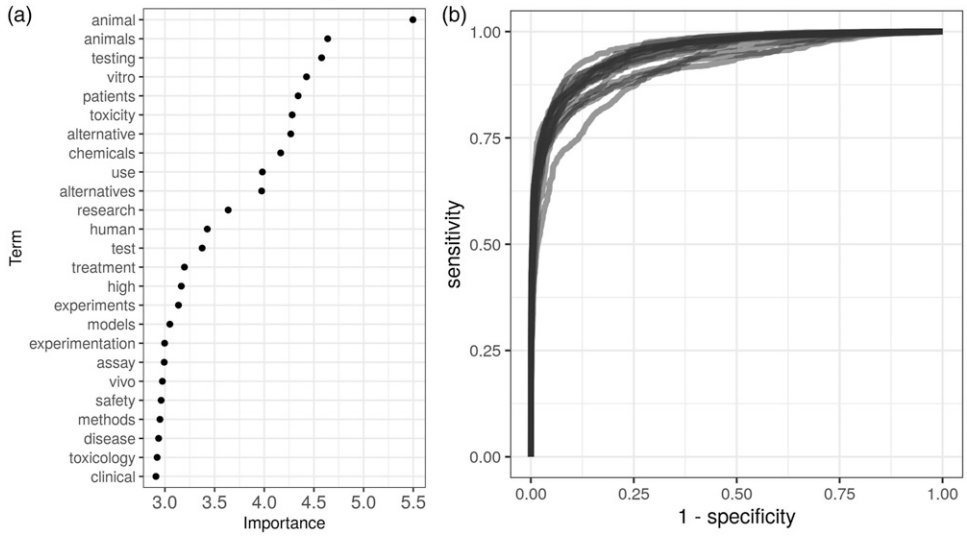

Figure 1. The generation of the *de novo* model with Random Forest. (a) Shows the most predictive features, as identified by the importance score of the Random Forest model. (b) Shows the Receiver Operator Characteristics (ROC) curves for 20 models.

Table 1. The titles of the five highest scoring papers for the 20,000 most recent dermatitis abstracts and the 20,000 most recent Parkinson's disease abstracts, scored with the *de novo* model.

Title	Score	Abstract set	Reference
Exposure of α-synuclein aggregates to organotypic slice cultures recapitulates key molecular features of Parkinson's disease	0.982	Parkinson's disease	Moudio et al.[19]
Avoiding contact allergens: From basic research to the *in vitro* identification of contact allergens	0.978	dermatitis	Martin and Esser[31]
Enhanced tyrosine hydroxylase activity induces oxidative stress, causes accumulation of autotoxic catecholamine metabolites, and augments amphetamine effects *in vivo*	0.976	Parkinson's disease	Vecchio et al.[20]
Adjustment of a no expected sensitization induction level derived from Bayesian network integrated testing strategy for skin sensitization risk assessment	0.976	dermatitis	Otsubo et al.[32]
The activity of methacrylate esters in skin sensitisation test methods II. A review of complementary and additional analyses	0.970	dermatitis	Kimber[33]
Computational modelling and simulation for immunotoxicity prediction induced by skin sensitisers	0.968	dermatitis	Russo et al.[34]
Ethosomes for dermal administration of natural active molecules	0.968	dermatitis	Natsheh et al.[35]
Evaluation of a low-toxicity PARP inhibitor as a neuroprotective agent for Parkinson's disease	0.958	Parkinson's disease	Puentes et al.[22]
Daphne genkwa flower extract promotes the neuroprotective effects of microglia	0.956	Parkinson's disease	Gupta et al. [23]
LncRNA XIST sponges miR-199a-3p to modulate the Sp1/LRRK2 signal pathway to accelerate Parkinson's disease progression	0.956	Parkinson's disease	Zhou et al.[21]

For each set, the five highest scoring abstracts were evaluated. In the dermatitis set, four out of the five papers did indeed describe non-animal alternatives, whereas one paper was a review that included both animal and non-animal methods. For the abstracts obtained with the Parkinson's disease search term, the observations were mixed. For example:

- Moudio et al.[19] describe the use of *ex vivo* organotypic brain slice cultures and suggest that these can be used to reduce animal experimentation for Parkinson's disease;[19,20] they also mention the results of *in vitro* experiments which they seek to corroborate with *in vivo* experiments;[20]
- Zhou et al.[21] actively compared the results of *in vitro* and *in vivo* models to examine the effect of lncRNA on Sp1 expression and LLRK2 functioning;[21]
- Puentes et al.[22] and Gupta et al.[23] both describe *in vitro* models, with the former suggesting validation of their

results in animal studies, the latter performing animal studies to corroborate their *in vitro* findings.

Taken together, these results indicate that the *de novo* model is very capable of identifying papers that feature potential Three Rs alternatives, out of a large set of abstracts returned by a PubMed query. However, as exemplified by the results obtained for the abstracts related to Parkinson's disease, there is still an opportunity for finetuning and increasing the performance of the model for this domain.

Improvement of the de novo *model*

To improve the performance of the *de novo* model on domain-specific papers with a focus on *replacement* (rather than *refinement* and *reduction*), the abstract sets used for training the model were filtered by using the assessments of

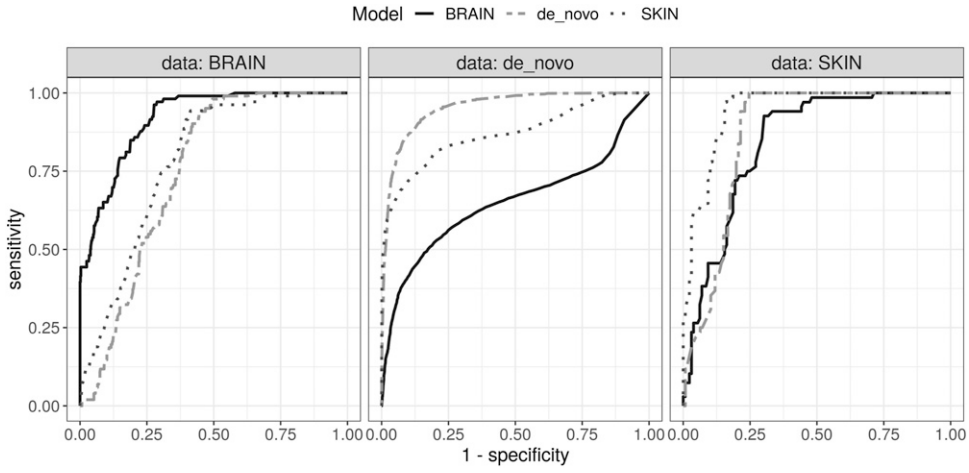

Figure 2. Performance of the BRAIN, *de novo* and SKIN models on the respective datasets, expressed as Receiver Operator Characteristics (ROC) curves. The datasets that were used to generate the ROC curves for the respective models are indicated in the panels above the graphs.

Table 2. Titles and references of the five highest scoring papers for the 20,000 most recent dermatitis abstracts and the 20,000 most recent Parkinson's disease abstracts, scored with the SKIN and BRAIN models.

Title	Score	Abstract set[a]	Reference
Utilising induced pluripotent stem cells in neurodegenerative disease research: Focus on glia	0.858	Parkinson's disease	Albert et al.[36]
Modeling neurodegenerative diseases with cerebral organoids and other three-dimensional culture systems: Focus on Alzheimer's disease	0.774	Parkinson's disease	Venkataraman et al.[37]
Therapeutic potential of pluripotent stem cell-derived dopaminergic progenitors in Parkinson's disease: A systematic review protocol	0.772	Parkinson's disease	Karimi et al.[24]
Unannotated single nucleotide polymorphisms in the TATA box of erythropoiesis genes show *in vitro* positive involvements in cognitive and mental disorders	0.742	Parkinson's disease	Ponomarenko et al.[38]
Maackiain ameliorates 6-hydroxydopamine and SNCA pathologies by modulating the PINK1/Parkin pathway in models of Parkinson's disease in *Caenorhabditis elegans* and the SH-SY5Y cell line	0.718	Parkinson's disease	Tsai et al.[39]
Avoiding contact allergens: From basic research to the *in vitro* identification of contact allergens	0.690	dermatitis	Martin and Esser[31]
Alternative methods for skin-sensitization assessment	0.678	dermatitis	Gądarowska et al.[40]
The activity of methacrylate esters in skin sensitisation test methods II. A review of complementary and additional analyses	0.670	dermatitis	Kimber[33]
Immune-competent *in vitro* co-culture models as an approach for skin sensitisation assessment	0.664	dermatitis	Th'elu et al.[41]
In silico prediction of skin sensitization: Quo vadis?	0.656	dermatitis	Ta et al.[42]

[a]Dermatitis abstracts were scored with the SKIN model; Parkinson's disease abstracts were scored with the BRAIN model.

the external reviewers, as described in the *Materials and methods* section. These experts scored each paper in accordance with definition of *replacement* in the EU *Directive 2010/63EU*: *"A scientifically satisfactory method or testing strategy, not entailing the use of live animals, shall be used instead of a procedure. This implies any scientific method employing non-sentient material replacing or avoiding using conscious whole living animals in experiments where they would otherwise have been used."*

We decided to focus on one model for skin research (the SKIN model) and one model for neurodegenerative diseases research (the BRAIN model). To this end, two sets of 20,000 abstracts were created, with the queries *skin disease* and *Parkinson's disease OR Alzheimer's disease OR frontotemporal dementia*. All abstracts were scored with the *de novo* model. Based on the scores attributed by the model, the abstracts were divided into five score classes: 0–0.2; 0.2–0.4; 0.4–0.6; 0.6–0.8; and 0.8–1. Then, 200 random abstracts were sampled from each score class, which resulted in a set of 1000 abstracts that were presented to the reviewers, without them having prior knowledge of the score assigned by the model. Based on the reviewer-curated sets of abstracts, two new RF models were created, one for each dataset (i.e. SKIN and BRAIN). To assess the performance of all three models — *de novo*, SKIN and BRAIN — the AUCs of all the models were measured on all the datasets. The AUCs show the trade-off between sensitivity and specificity for the three models, at all possible thresholds for the probability (Figure 2).

The results showed that all models have the best performance on the datasets that were used for training the specific model, with AUC values ranging from 0.954 for the SKIN model, to 0.91 for the BRAIN model. In addition, the models also exhibit good performance on the other datasets, ranging from 0.82 (SKIN model on the BRAIN dataset) to 0.65 (BRAIN model on the *de novo* dataset).

Application of the SKIN and BRAIN models

To investigate how the models would work in practice, the SKIN and BRAIN models were used to score the same set of abstracts for dermatitis and Parkinson's disease, respectively, to assess the performance of the *de novo* model. The top five papers from both sets of abstracts were inspected manually and are shown in Table 2.

The data show that, when using the BRAIN model, a number of papers were identified that clearly describe new *in vitro* methods — in this case mostly based on the

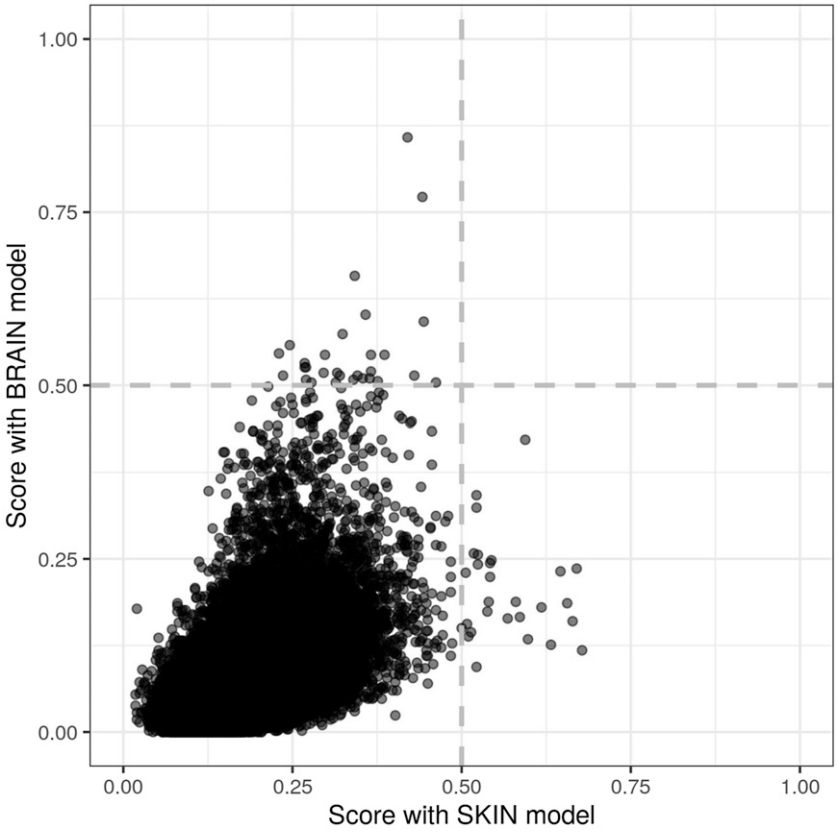

Figure 3. A comparison of the abstract scores by the SKIN model and the BRAIN model. Each point reflects an individual abstract that has been scored separately by both models, the scores of which are indicated on the x-axis and y-axis.

use of pluripotent stem cells. It should be noted that the paper by Karimi et al.,[24] which is a systematic review, also includes the use of pluripotent stem cells in preclinical animal models. Similarly, when using the SKIN model, papers that explicitly describe *in vitro* methods that could serve as replacement methods were identified.

For a further validation process, the top 200 abstracts based on the 3Ranker score (the '3Ranker set') and the top 200 abstracts based on the default PubMed ordering (i.e. recent papers first; the 'PubMed set') were selected, from both the SKIN and the BRAIN datasets. These 800 abstracts were anonymised and presented in random order, to six independent reviewers (three for the SKIN dataset and three for the BRAIN dataset) who had not been involved in the model building process. These abstracts were then manually screened and classified according to their Three Rs relevance, i.e. Three Rs or non-Three Rs. The numbers of papers manually classified as Three Rs papers in the 3Ranker set and in the PubMed set were then compared. For the SKIN dataset, the reviewers found an average 8.6-fold enrichment of Three Rs papers in the 3Ranker set, as compared to the PubMed set. For the BRAIN dataset, a 10.3-fold enrichment for the 3Ranker set, as compared to the PubMed set, was found. In classifying the abstracts, there was 86% manual reviewer agreement for the SKIN dataset and 81% manual reviewer agreement for the BRAIN dataset.

Comparison of the BRAIN model versus the SKIN model

To investigate whether the BRAIN and SKIN models really do possess domain-specific adaptation, any potential overlap in the predictive power of both models was investigated. To this end, abstracts for 'dermatitis' and for 'Parkinson's disease' were scored with both models; the results of this scoring are shown in Figure 3.

There appears to be a weak correlation between the two scores obtained for each individual abstract with both models, but it is clear that not all of the abstracts highly-scored by the SKIN model are also scored highly by the BRAIN model, and vice versa. This indicates that there is some predictive overlap between the base models, but we expect that this could be mitigated by a process of domain-specific adaptation to improve the model further.

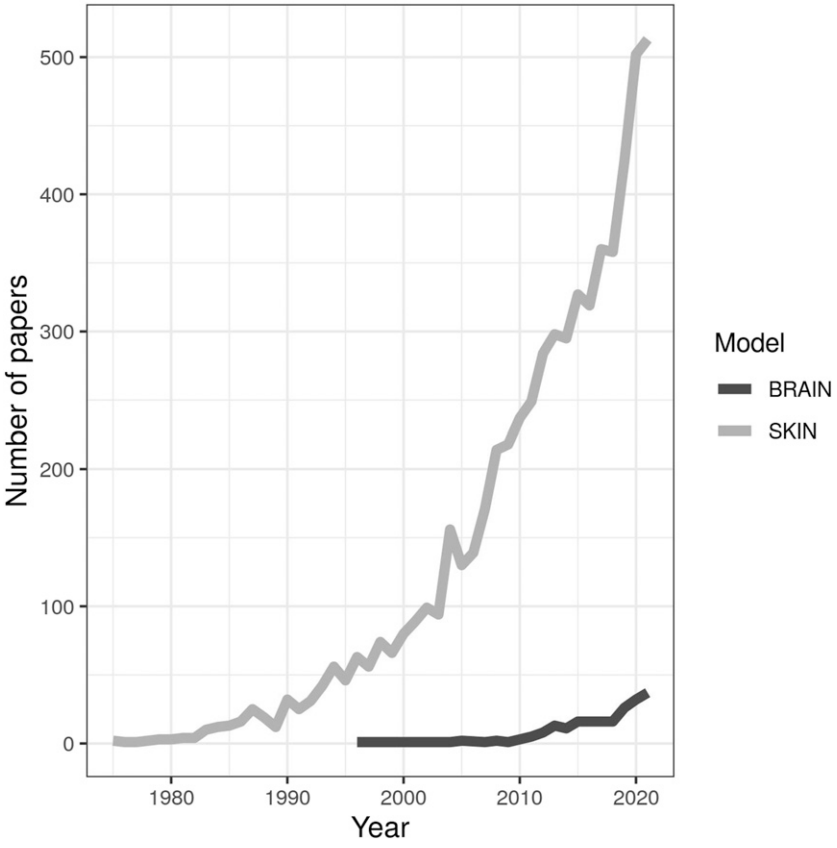

Figure 4. The number of MEDLINE abstracts with a positive score (> 0.5) for the SKIN or the BRAIN model, calculated per year.

Tracking Three Rs papers over time

A potential interesting application of the models is their use to track the number of publications on the Three Rs over time. To investigate how the number of Three Rs publications has changed over time, we used both SKIN and BRAIN models to score all 35,073,619 MEDLINE abstracts. For each model, the number of abstracts that scored above a threshold of 0.5 was counted and grouped by year.

The number of abstracts with a high score, and their rate of increase over time, was greater with the SKIN model than with the BRAIN model (Figure 4), indicating that progress in the area of replacement strategies varies across disciplines. This may be due to differences in subject-specific obstacles that need to be overcome, or societal/governmental influence on the use (or otherwise) of specific testing and conditions. For example, an important contributory factor to the higher number/rate of increase of abstracts identified with the SKIN model could be the legislative ban on animal testing for cosmetics in the EU (*Regulation (EC) No 1223/2009*),[25] which focused attention on the development of Three Rs alternative test methods in the field.

Availability

To enable researchers to make use of these models, a web-based search interface was created and made available at www.open3r.org (Figure 5). Through this interface, users can search with any keyword as a search string across all abstracts indexed in MEDLINE. The search results are then scored by using both the BRAIN and SKIN model, and the 200 papers with the highest score for each model are displayed in a results table. In addition to the search function, the abstracts that were identified as being related to the Three Rs (specifically, *replacement*) by the expert reviewers, and thus used as positive indicators in the RF classifiers, are shown under the 'Curated Abstracts' tab. This table can be filtered according to the subject of interest, by using the search box in the top-right corner. The content of this table is iteratively updated after new active-review rounds of the 3Ranker algorithm. The curation process is facilitated by the 'Reference Scoring' feature. This is an

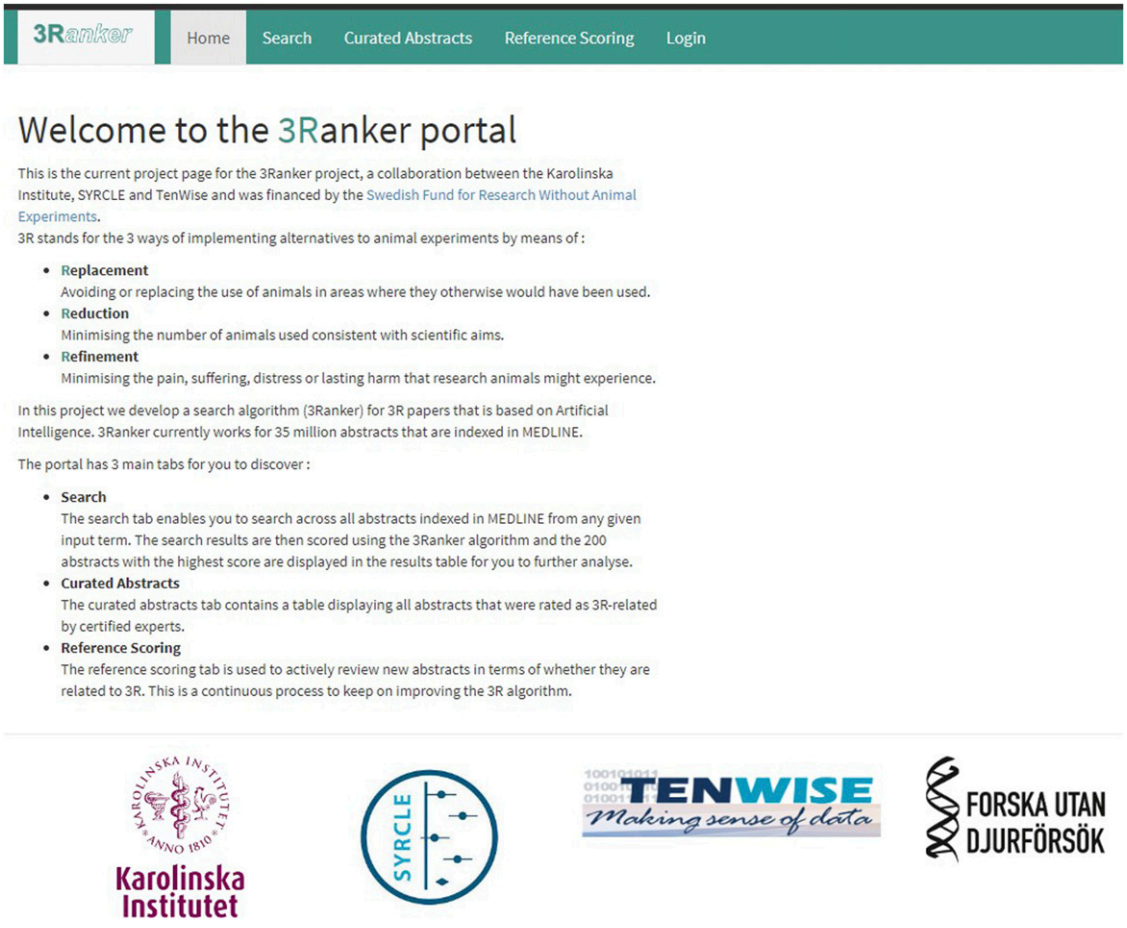

Figure 5. A screenshot of the www.open3r.org web application. Please see the main text for a description of the various features.

active learning modus that is used by the expert reviewers to actively review new abstracts in terms of whether they are related to the Three Rs. These review rounds are iteratively scheduled to design and create new classifiers for Three Rs-related abstracts, which are then added to the 'Search Mode' feature.

Conclusions and future work

The results presented in this paper show that it is possible to create an AI-based classification model capable of rapidly prioritising biomedical research papers with respect to their relevance to the Three Rs. The added value of our work over previously published classification algorithms is the fact that the development of the model has been directly coupled to the indexed MEDLINE database, and it is also a freely-available web-based application. This provides users with a natural starting point (a general PubMed query), combined with a direct link to the ranked results. Various commercial and publicly available software for managing systematic reviews, such as DistillerSR, Eppi Reviewer, Abstrackr,[26] employ classification models for the computer-assisted screening of abstracts. However, these tools require user registration and the need to preload the system with abstracts. In addition, the AI models that are created on the basis of the screening results are only available within the specific screening projects. In practice, this prevents the widespread use of these tools in searches for non-animal replacements.

We showed that even a model based on a starting set created with weak supervision — i.e. a set of abstracts that was enriched for Three Rs papers but not yet curated by experts — could yield good results. However, a considerable improvement of this starting model toward a domain-specific model could be made by involving human experts, which indicates that a combined approach of machine learning and human expert input yields the best results. This corroborates the findings described elsewhere.[9] The current improvements to the starting model were aimed at creating two separate models for skin and brain research, and we foresee that similar improvements could be made to establish models for other diseases, such as cancer and cardiovascular diseases.

In the current project, the abstracts were all classified as being related to the Three Rs — which is, to some extent, a blanket term. Therefore, in addition to diversification of the model into specific research domains, further improvements should aim to create a model that more precisely describes the specific nature of the Three Rs content. This should include the ability to delineate whether the paper features a *replacement*, *reduction* or *refinement* approach, as well as the type of alternative method described — for example, whether it is an *in silico* method, *in vitro* method, read-across approach, or organ-on-a-chip application.[27]

In this paper we used RF modelling in combination with a 'Bag-of-Words' (BOW) as input data, to make a generalised tool that can also be used to classify text from various other sources. A BOW approach is based on the number of occurrences of a particular word in an excerpt of text. Although this approach may yield models that perform well, BOW methods have a number of limitations — for example, they cannot handle multi-word phrases and do not reflect the variable importance of different parts of the text. Models that are trained with a BOW as the input data do not take into account the semantic meaning of the words, or the structure of the sentences. Artificial intelligence models that use word and paragraph embedding are likely to perform better on these kinds of tasks.[5,28–30] Future work could therefore focus on the use of these types of dedicated models for classifying papers on the Three Rs. Given that we have an end-to-end framework, in which training and application of the model for the user are combined in a single platform selection, new models from the fast-evolving field of language-based AI models can be rapidly deployed for use.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants F2021-0011 and N2020-0009 from the Swedish Foundation for Replacement in Animal Research.

ORCID iDs

Sinne Alkema https://orcid.org/0000-0002-2463-5196
Brun Ulfhake https://orcid.org/0000-0001-5037-7461
Merel Ritskes-Hoitinga https://orcid.org/0000-0001-5315-284X
Wynand Alkema https://orcid.org/0000-0002-5523-9217

References

1. van Luijk J, Leenaars M, Hooijmans C, et al. Towards evidence-based translational research: The pros and cons of conducting systematic reviews of animal studies. *ALTEX* 2013; 30: 256–257.
2. Ratajeski MA and Miller RS. Adapting to changes in publishing when searching for alternatives and reporting on animal research: A librarian's perspective. *Altern Lab Anim* 2022; 50: 57–61.
3. de Vries RB, Hooijmans CR, Tillema A, et al. A search filter for increasing the retrieval of animal studies in Embase. *Lab Anim* 2011; 45: 268–270.

4. Yauw STK, Wever KE, Hoesseini A, et al. Systematic review of experimental studies on intestinal anastomosis. *Br J Surg* 2015; 102: 726–734.
5. Habets PC, van IJzendoorn DG, Vinkers CH, et al. Development and validation of a machine-learning algorithm to predict the relevance of scientific articles within the field of teratology. *Reprod Toxicol* 2022; 113: 150–154.
6. Wang Y, Sohn S, Liu S, et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 2019; 19: 1.
7. Fischer I and Steiger H-J. Toward automatic evaluation of medical abstracts: The current value of sentiment analysis and machine learning for classification of the importance of PubMed abstracts of randomized trials for stroke. *J Stroke Cerebrovasc Dis* 2020; 29: 105042.
8. Marshall IJ, Noel-Storr A, Kuiper J, et al. Machine learning for identifying randomized controlled trials: An evaluation and practitioner's guide. *Res Synth Methods* 2018; 9: 602–614.
9. Bannach-Brown A, Przybyła P, Thomas J, et al. Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error. *Syst Rev* 2019; 8: 23.
10. Ritskes-Hoitinga M and Alkema W. The use of artificial intelligence for the fast and effective identification of three Rs-based literature. *Altern Lab Anim* 2021; 49: 133–136.
11. Speit G. How to assess the mutagenic potential of cosmetic products without animal tests? *Mutat Res* 2009; 678: 108–112.
12. Johansson H and Lindstedt M. Prediction of skin sensitizers using alternative methods to animal experimentation. *Basic Clin Pharmacol Toxicol* 2014; 115: 110–117.
13. Fentem JH. The 19th FRAME Annual Lecture, November 2022: Safer chemicals and sustainable innovation will be achieved by regulatory use of modern safety science, not by more animal testing. *Altern Lab Anim* 2023; 51: 90–101.
14. Cassotta M, Geerts H, Harbom L, et al. The future of Parkinson's disease research: A new paradigm of human-specific investigation is necessary… and possible. *ALTEX* 2022; 39: 694–709.
15. Feinerer I, Hornik K and Meyer D. Text mining infrastructure in R. *J Stat Softw* 2008; 25: 1–54.
16. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Softw* 2019; 4: 1686.
17. Liaw A and Wiener M. Classification and regression by RandomForest. *R News* 2002; 2: 18–22.
18. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; 12: 77.
19. Moudio S, Rodin F, Albargothy NJ, et al. Exposure of α-synuclein aggregates to organotypic slice cultures recapitulates key molecular features of Parkinson's disease. *Front Neurol* 2022; 13: 826102.
20. Vecchio LM, Sullivan P, Dunn AR, et al. Enhanced tyrosine hydroxylase activity induces oxidative stress, causes accumulation of autotoxic catecholamine metabolites, and augments amphetamine effects *in vivo*. *J Neurochem* 2021; 158: 960–979.
21. Zhou Q, Zhang M-M, Liu M, et al. LncRNA XIST sponges miR-199a-3p to modulate the Sp1/LRRK2 signal pathway to accelerate Parkinson's disease progression. *Aging* 2021; 13: 4115–4137.
22. Puentes LN, Lengyel-Zhand Z, Reilly SW, et al. Evaluation of a low-toxicity PARP inhibitor as a neuroprotective agent for Parkinson's disease. *Mol Neurobiol* 2021; 58: 3641–3652.
23. Gupta DP, Park SH, Lee Y-S, et al. *Daphne genkwa* flower extract promotes the neuroprotective effects of microglia. *Phytomedicine* 2023; 108: 154486.
24. Karimi A, Elmi M, Shiri Z, et al. Therapeutic potential of pluripotent stem cell-derived dopaminergic progenitors in Parkinson disease: A systematic review protocol. *Syst Rev* 2021; 10: 188.
25. European Union. *Regulation (EC) No 1223/2009* of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Off J Eur Union* 2009; L342: 59–209.
26. van der Mierden S, Tsaioun K, Bleich A, et al. Software tools for literature screening in systematic reviews in biomedical research. *ALTEX* 2019; 36: 508–517.
27. Gilmour N, Reynolds J, Przybylak K, et al. Next generation risk assessment for skin allergy: Decision making using new approach methodologies. *Regul Toxicol Pharmacol* 2022; 131: 105159.
28. Shen Z, Schutte D, Yi Y, et al. Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. *BMC Med Inform Decis Mak* 2022; 22: 88.
29. Rabby G and Berka P. Multi-class classification of COVID-19 documents using machine learning algorithms. *J Intell Inf Syst* 2023; 60: 571–591.
30. Del Fiol G, Michelson M, Iorio A, et al. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: Comparative analytic study. *J Med Internet Res* 2018; 20: e10281.
31. Martin SF and Esser PR. Avoiding contact allergens: From basic research to the *in vitro* identification of contact allergens. *Allergol Select* 2017; 1: 77–84.
32. Otsubo Y, Nishijo T, Mizumachi H, et al. Adjustment of a no expected sensitization induction level derived from Bayesian network integrated testing strategy for skin sensitization risk assessment. *J Toxicol Sci* 2020; 45: 57–67.
33. Kimber I. The activity of methacrylate esters in skin sensitisation test methods II. A review of complementary and additional analyses. *Regul Toxicol Pharmacol* 2021; 119: 104821.

34. Russo G, Crispino E, Corsini E, et al. Computational modelling and simulation for immunotoxicity prediction induced by skin sensitisers. *Comput Struct Biotechnol J* 2022; 20: 6172–6181.
35. Natsheh H, Vettorato E and Touitou E. Ethosomes for dermal administration of natural active molecules. *Curr Pharm Des* 2019; 25: 2338–2348.
36. Albert K, Niskanen J, Kälvälä S, et al. Utilising induced pluripotent stem cells in neurodegenerative disease research: Focus on glia. *Internat J Mol Sci* 2021; 22: 4334.
37. Venkataraman L, Fair SR, McElroy CA, et al. Modeling neurodegenerative diseases with cerebral organoids and other three-dimensional culture systems: Focus on Alzheimer's disease. *Stem Cell Rev Rep* 2022; 18: 696–717.
38. Ponomarenko M, Sharypova E, Drachkova I, et al. Unannotated single nucleotide polymorphisms in the TATA box of erythropoiesis genes show *in vitro* positive involvements in cognitive and mental disorders. *BMC Med Genet* 2020; 21 (Suppl. 1): 165.
39. Tsai R-T, Tsai C-W, Liu S-P, et al. Maackiain ameliorates 6-hydroxydopamine and SNCA pathologies by modulating the PINK1/parkin pathway in models of Parkinson's disease in *Caenorhabditis elegans* and the SH-SY5Y cell line. *Int J Mol Sci* 2020; 21: 4455.
40. Gądarowska D, Kalka J, Daniel-Wójcik A, et al. Alternative methods for skin-sensitization assessment. *Toxics* 2022; 10: 740.
41. Th'elu A, Catoire S and Kerdine-Römer S. Immune-competent *in vitro* co-culture models as an approach for skin sensitisation assessment. *Toxicol In Vitro* 2020; 62: 104691.
42. Ta GH, Weng C-F and Leong MK. *In silico* prediction of skin sensitization: Quo vadis? *Front Pharmacol* 2021; 12: 655771.