AI-aided Systematic Review to Create a Database with Potentially Relevant Papers on Depression, Anxiety, and Addiction

Authors: Marlies Brouwer[1*], Laura Hofstee[2], Sofie van den Brand[2], Jelle Teijema[2], Gerbrich Ferdinands[3], Jan de Boer[3], Felix Weijdema[3], Bianca Kramer[3], Reinout Wiers[4], Claudi Bockting[1], Jonathan de Bruin[5], Rens van de Schoot[2]

[1] Amsterdam UMC, location University of Amsterdam, Department of Psychiatry and Centre for Urban Mental Health, University of Amsterdam, The Netherlands

[2] Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, The Netherlands

[3] Utrecht University Library, Utrecht University, The Netherlands

[4] Centre for Urban Mental Health and department of psychology UvA

[5] Department of Research and Data Management Services, Information Technology Services, Utrecht University, Utrecht, the Netherlands

Corresponding author: Marlies Brouwer, Amsterdam UMC, location University of Amsterdam, Department of Psychiatry, Centre for Urban Mental Health, m.e.brouwer@amsterdamumc.nl

Funding: This project is funded by a grant from the Centre for Urban Mental Health, University of Amsterdam, The Netherlands.

Acknowledgment: We would like to thank all involved researchers and assistants in this project for their input, assistance, and advices: Prof. P. Lucassen, Prof. P. Sloot, Prof. K. Stronks, Prof. J. van Weert, senior researchers from Centre for Urban Mental Health (UMH), Dr. V. Melnikov, P. Schulte-Frankenfeld, BA, Veronica Szpak, MSc, Matthew Salanitro, MSc, Daniele Franzoi, MSc, Linda Betelli, MSc, Emanuela Zhecheva, BSc, Milena Kapralova, BSc, Melissa Dorrestein, BSc, Annel Koomen, MSc, Lucie Pressl, BSc, Mona Weingaertner, BSc, Beau Manancourt, BSc, and Nitya Shah, BSc.

Abstract

It is of utmost importance to provide an overview and strength of evidence of predictive factors and to investigate the current state of affairs on evidence for all published and hypothesized factors that contribute to the onset, relapse, and maintenance of anxiety-, substance use-, and depressive disorders. Thousands of such articles have been published on potential factors of CMDs, yet a clear overview of all preceding factors and interaction between factors is missing. Therefore, the main aim of the current project was to create a database with potentially relevant papers obtained via a systematic. The current paper describes every step of the process of constructing the database, from search query to database. After a broad search and cleaning of the data, we used active learning using a shallow classifier and labeled the first set of papers. Then, we applied a second screening phase in which we switched to a different active learning model (i.e., a neural net) to identify difficult-to-find papers due to concept ambiguity. In the third round of screening, we checked for incorrectly included/excluded papers in a quality assessment procedure resulting in the final database. All scripts, data files, and output files of the software are available via Zenodo (for Github code), the Open Science Framework (for protocols, output), and DANS (for the datasets) and are referred to in the specific sections, thereby making the project fully reproducible.

Introduction

Common mental conditions, including anxiety, substance use, and depression, are prevalent and disabling conditions, affecting an estimated 20% of people globally each year (Steel et al., 2014). Prevalence rates are even higher when including harmful use of substances such as tobacco and alcohol (Effertz & Mann, 2013). These common mental disorders (CMDs) cause a burden on individuals, their family and friends, and the nations, pose a high risk of suicidality, decrease overall life expectancy, and cause a lower overall quality of life. For this reason, it is of utmost importance to provide an overview of predictive factors, including the strength of their evidence, that contribute to the onset, relapse, and maintenance of anxiety-, substance use-, and depressive disorders. Thousands of such articles have been published on potential factors of CMDs, yet a clear overview of all preceding factors and interaction between factors is missing.

Therefore, the main aim of the current project was to create a database with potentially relevant papers obtained via a systematic literature search. Developing a search strategy for a systematic review is an iterative process to balance recall and precision (Lefebvre et al., 2008). That is, including as many potentially relevant studies as possible (recall) while simultaneously limiting the total number of studies (precision). Given that screening the entire research literature on a given topic is too labor-intensive (Borah et al., 2017), scholars often develop narrower searches, with the risk of missing relevant studies and potentially neglecting research areas relevant to CMDs. Performing a systematic search is time-consuming, and it is a tedious task to obtain an overview of the entire field of CMDs. Systematic literature searches result in the identification of millions of published manuscripts that claim to have studies factors of CMDs, yet a clear overview of all preceding factors and interaction between factors is missing. Indeed, for a recent meta-analysis on the evidence for leading psychological and biological theories on the onset, maintenance, and relapse of depressive disorders, almost 150.000 records needed to be screened, which took three years with a team of 18 researchers (Brouwer et al., 2019; Fu et al., 2021; Kennis et al., 2020).

To deal with the enormous amount of papers identified in a search, machine learning can assist in finding relevant texts much faster (Harrison et al., 2020). A well-established approach to increasing the efficiency of title and abstract screening is screening prioritization (Cohen et al., 2009) with active learning (Settles, 2012). With active learning-based systematic reviewing, the labeling actions of humans are used to train a new model which selects the most likely relevant paper from the set of unseen papers. The primary output of the process is a selection of the relevant papers, not the model itself, with a minimum number of labeling tasks. Simulation studies show that machine-learning-based prioritization with active learning enables us to find relevant studies much faster than traditional screening methods; it can save up to 95% of screening time (Van de Schoot, De Bruin, Schram, Zahedi, De Boer, Weijdema, Kramer, Huijts, Ferdinands, Harkema, Harkema, Willemsen, Ma, Fang, Sybren, et al., 2021). For example, a simulation study was performed on labeled data retrieved from a previously described meta-analysis (Brouwer et al., 2019) on the prospective evidence for leading psychological theories of depressive relapse (Teijema et al., 2022). The total number of papers found in the search was 50.936, of which 63 were included in the final meta-analysis (excluding the papers found with snowballing). The

simulation study performed on this data indicated that after reading only 10% of the papers using active learning, 62 of the 63 papers were already found. When re-analyzing the meta-analytic results, it appeared that the conclusions would have been the same when excluding the one paper missed. Even when removing the 10% most difficult to find paper, the overall results remained the same for the analyses on time to depressive relapse (Hazard Ratio).

Machine-learning prioritization with active learning aims to save time by screening fewer papers than exist in the entire pool. This saves valuable time or broadens the scope of the search and includes a larger pool of papers, so results depend less on the quality of the initial search process (Gusenbauer & Haddaway, 2020). The innovative use of active learning allows for broader use of the pipeline by omitting limitations in the search query, i.e., attaining a broad search query. A broader search query results in a higher number of initial papers without replacing the initial collection step and limits the risk of missing relevant papers (i.e., better recall). With the same time investment, many more papers can be included in the screening process, which is especially relevant for our current study. There can be a risk of missing relevant papers in the search query due to different terminology, which may result in missing crucial evidence.

To construct a database on the topic of CMDs, we plan to conduct a comprehensive search query, probably resulting in a dataset too big for humans to screen in a limited timeframe. Instead, we plan to apply three rounds of screening using active learning to identify as many potentially relevant papers as possible while balancing screening time. The current paper describes every step of the process of constructing the database, from search query to the final database. All scripts, data files, output files of the software are available via Zenodo (for Github code), the Open Science Framework (for protocols, output), and DANS (for the datasets; Brouwer (2022b)) and are referred to in the specific sections, thereby making the project fully reproducible. The final data is available on the Dutch national centre of expertise and repository for research data DANS (Brouwer, 2022a).

Methods

Search and pre-processing data

To build the search strategy, we used previous review articles, and the authors received input from various experts in the field of CMDs. A list of relevant search terms related to the disorders and potential predictive factors were created. For example, this included terms such as ‘Major depressive disorder’, ‘panic’, ‘predictor’, and ‘diabetes’. Before the actual search was conducted, the searches were first tested and discussed with an information specialist and the research group. The final search was conducted in June 2021, and papers were identified via a search carried out in four search engines (Embase, Medline, PsycINFO, and Scopus) for three different topics (anxiety, depressive-, and substance use disorders). The imported identified records from the 4 databases for each disorder were combined into a separate topic dataset in Endnote. For each topic dataset, we first used the option to automatically update references in Endnote X9 to

find missing abstracts and digital object identifiers (DOIs). Then, we used the automatic deduplication steps as described in (Bramer et al., 2016). Next, since Endnote does not offer deduplication based on DOI, we created a script in R. The set search terms, the search queries, and pre-processing steps are published on the Open Science Framework (Brouwer et al., 2021). The output of the search (i.e,. dataset) is available on DANS (Brouwer, 2022b).

Screening Phase 1: Active Learning

We used the software ASReview version 0.17 (Van de Schoot, De Bruin, Schram, Zahedi, De Boer, Weijdema, Kramer, Huijts, Ferdinands, Harkema, Harkema, Willemsen, Ma, Fang, Tummers, et al., 2021) installed on a server. The virtual machine could only be accessed from within the university network via a secured VPN connection. The server installation, security, backup facilities, and installation steps are described in (Melnikov, 2021).

As training data, a set of key papers were selected about each topic (anxiety, depressive-, and substance use disorders) to represent a wide variety of related topics (i.e., behavior, biological, demographic, psychological, urban, treatment change, resistant/maintenance, onset, relapse/recurrence, and recovery). The set of key papers per topic is available at the OSF (Brouwer et al., 2021). The selected key papers were used to train the first iteration of the active learning model. In the first screening phase, we used logistic regression as the classifier and TF-IDF as the feature extractor. The settings of the active learning model were based on a simulation study performed by Teijema et al. (2022) on the depression dataset of Brouwer et al. (2019).

The five screeners were instructed to screen records using a decision tree, see Fig. 1. If needed, the screeners could screen the full record to decide on inclusion. The screeners could not screen the same project – dataset – simultaneously; the screeners were locked out of the project when another screener accessed it. They were asked to screen for one session maximum of four hours before changing the topic to avoid screening fatigue. If the decision could not be made using the title/abstract, the screeners were instructed to search for the full-text and make the full-text decision directly in the software. It could, in theory, happen that one of the screeners labeled the first duplicate paper to be relevant, stopped the screening session (at the end of a session), and another screener (starting a new session) labeled the second duplicate paper to be irrelevant. A protocol was developed to prevent applying noisy labels on how to end a screening session (Hofstee et al., 2021). The stopping rule of the first screening phase was a two-fold rule; screening in this phase was stopped when more than 90 percent of irrelevant papers in a batch of 100 papers – a plateau – were found or when the time had run out of the screeners (i.e., 220 *contract hours*).

The three export files from ASReview containing all labeling decisions of the active learning process for each topic, as well as the input data (i.e., meta-data of the papers), plus the labeling decisions, are available via DANS (Brouwer, 2022b).

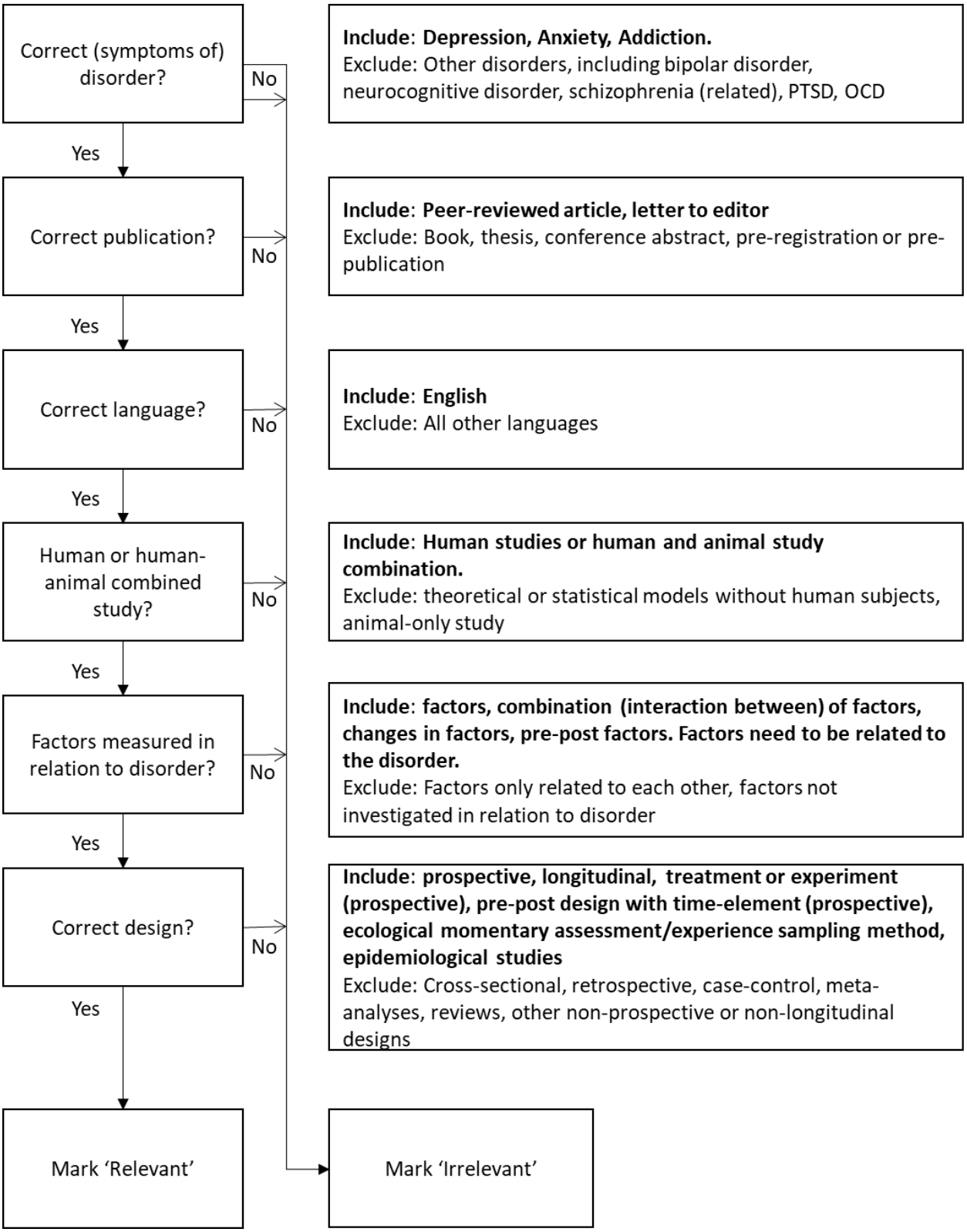

Fig 1. Decision tree for the first screening phase.

Screening phase 2: Deep Learning

The goal for the second screening phase was to switch to a different active learning model. That is, it might be that papers can be difficult to find due to concept ambiguity (i.e., concept drift) (Chen et al., 2012; Gama et al., 2014), and the algorithms must “dig deeper” into a text to find its essence (Goodfellow et al., 2016). Deep learning networks, like neural networks, are better at finding complex connections within data when compared to shallow networks such as the logistic model we used for the first screening phase. However, such deep learning models require much more training data (Alwosheel et al., 2018), and a neural net is, in the first couple of iterations, not expected to perform well.

Therefore, for the second screening phase, we first used the labeling decisions of the first round to optimize the hyperparameters per topic to receive the optimal hyperparameters for a 17-layer convolutional neural network (CNN)(Teijema, 2021) in combination with a doc2vec feature extractor. This model appeared to have better performance than the default deep learning models available in ASReview as was concluded in a simulation study conducted on similar data (Teijema et al., 2022). Using the optimized hyperparameters, we trained the 17-layer CNN model. The optimization and training were done via a Jupyter Notebook run in Google Colab. All steps and scripts to reproduce these steps are described in (Teijema & Van de Schoot, 2021).

The three pre-trained models were applied to the unlabelled data and sent to three researchers (RvdS, JT, SvdB), whom each screened one topic for another two weeks up to a stopping rule of finding >90% irrelevant abstracts in a batch of 100 papers.

Again, the export files from ASReview containing all labeling decisions at each iteration of the active learning process and the input data (i.e., meta-data of the papers) plus the labeling decisions are available via DANS (Brouwer, 2022b).

Screening Phase 3: Quality checks

Merge data

After screening phase two was completed, the three datasets were combined into one overall dataset. A column was added indicating whether a paper was included in at least one of the topics. A deduplication strategy was applied since a paper could have been included in more than one of the datasets. To improve the deduplication process, missing DOIs were retrieved as much as possible with a Python notebook to find the missing DOIs with Crossref based on a title-year combination. Next, the dataset was deduplicated based on author, title, year, and journal/ISSN. Before deduplication, these columns were set to lowercase characters, and all punctuation marks were removed. The result of the deduplication strategy was used in the proceeding steps since it was assumed that it was better to miss duplicate papers than to deduplicate papers falsely. All scripts and much more detailed information about cleaning the data are available in van den Brand et al. (2021), and the resulting dataset is available on DANS (Brouwer, 2022b).

Quality check 1

To check for incorrectly excluded but relevant papers, we exported the papers for each subject area identified as ‘irrelevant’. We selected the 20 lowest-ranked irrelevant papers as training data (for each subject area), and we added the 20 highest-ranked relevant papers for each subject area as prior relevant records. An active learning model was trained on the three datasets using the default settings (Naive Bayes and TF-IDF). The first author screened each dataset (anxiety, depression, substance abuse) for about 2 hours to identify incorrectly excluded but relevant papers. The data with records to be included and ASReview project files are available on DANS (Brouwer, 2022b).

Quality check 2

MB and multiple screeners went through the included papers to check for incorrectly included but irrelevant papers. Irrelevant papers, such as reviews, qualitative studies, and retrospective studies, were marked and removed from the database. The data with records to be excluded are available on DANS (Brouwer, 2022b).

Final Database

After correcting the labels via an R-function as described in van den Brand et al. (2021). Table 1 indicates which columns (next to the original meta-data like title, author, and abstract) were created. The final database is available via DANS (Brouwer, 2022a).

Table 1. Overview of the column names generated in the final dataset.

Column name	Values	Description
index	1,…,N	A simple indexing column. Some numbers are not present because they have been removed after deduplication.
unique_paper	0, 1, NA	Indicating whether the column has a unique DOI. This is NA when there is no DOI present.
depression_included anxiety_included substance_included	0, 1, NA	A column indicating whether a paper was included per topic
quality_check_1(0->1)	1, 2, 3, NA	This column indicates for which subjects a paper was falsely excluded and relabelled (1=anxiety; 2=depression; 3=substance use)
quality_check_2(1->0)	1, 2, 3, NA	This column indicates for which subjects a paper was falsely included and relabelled (1=anxiety; 2=depression; 3=substance use)
depression_included_corrected substance_included_corrected anxiety_included_corrected	0, 1, NA	Combining the information from the screening phases and the quality checks, this column contains the inclusion/exclusion/not seen labels after correction.
composite_label_corrected	0, 1, NA	A column indicating whether a paper was included in at least one of the corrected_subject columns
Note: For all columns where there are only 0's 1's and NA's, a 0 indicates an irrelevant paper, while 1 indicates relevant, and NA means Not Available.		

Results

After searching four databases, initially 2,739,753 records were identified. Due to this large number of hits, adjustments of the search terms were necessary. After adjustments of the search terms, which are reported on DANS, 70,065 papers were found for anxiety disorders, 83.371 for substance use disorders, and 161.760 for depressive disorders, see the first row of the flow chart in Fig. 2. To illustrate how to read the flowchart, we use the anxiety data as an example.

The anxiety dataset was deduplicated, removing 36,086 duplicates, and reducing the total number of papers to 33,979. After expert consensus, 11 papers were labeled as relevant to be used as prior knowledge, and ten random records were read and labeled as irrelevant records. The partly labeled dataset (33,958 unlabelled and 21 labeled papers) was imported into the ASReview software, and an active learning model was trained. For the first screening phase, a team of 5 screeners supervised by MB and LH screened from June 21 to October 21, 2021 (17 weeks). Of the 2,729 papers that were screened, the team labeled 1,497 as relevant and 1,232 as irrelevant. The partly labeled dataset was used to train the hyperparameters of a 17-layer convolutional neural network (CNN) in combination with a doc2vec feature extractor (training time three days).

The pre-trained model was sent to the second team of screeners, whom each screened one (out of three) dataset for another two weeks. For the Anxiety dataset, 290 extra papers were screened and labeled (ir)relevant (relevant: 24; irrelevant: 266). For the Anxiety dataset, 100% of the last batch of 100 records screened were irrelevant (96% for the Substance use dataset; 95% for the Depression dataset).

Next, the three datasets were combined (total number of papers: 165,046) and again deduplicated (number of duplicate papers: 35,717), reducing the total number of papers to 129,329. For the first quality check, in total, 388 labels originally determined as irrelevant and predicted by the machine learning model as most likely relevant were assessed by MB, and 95 labels were converted to relevant. For the second quality check, in total 6.380 labels originally determined as relevant were checked, and 28 were converted to irrelevant.

Two databases were created:

(1) megameta_asreview_partly_labelled: A partly labeled database (n_relevant = 6,351; n_irrelevant= 4,411; n_unlabelled = 118,567) with the columns as described in Table 1. The data can be used for post-processing or can be imported into software to continue labeling using all previous labels as training data.

(2) megameta_asreview_only_potentially_relevant: A dataset with only the relevant papers (n= 6,380). This database can be used to search for papers for a specific research question. We also created a dataset per domain (n_anxiety = 1522; n_depression = 2708; n_substance = 2306). Note that some papers are relevant for multiple domains.

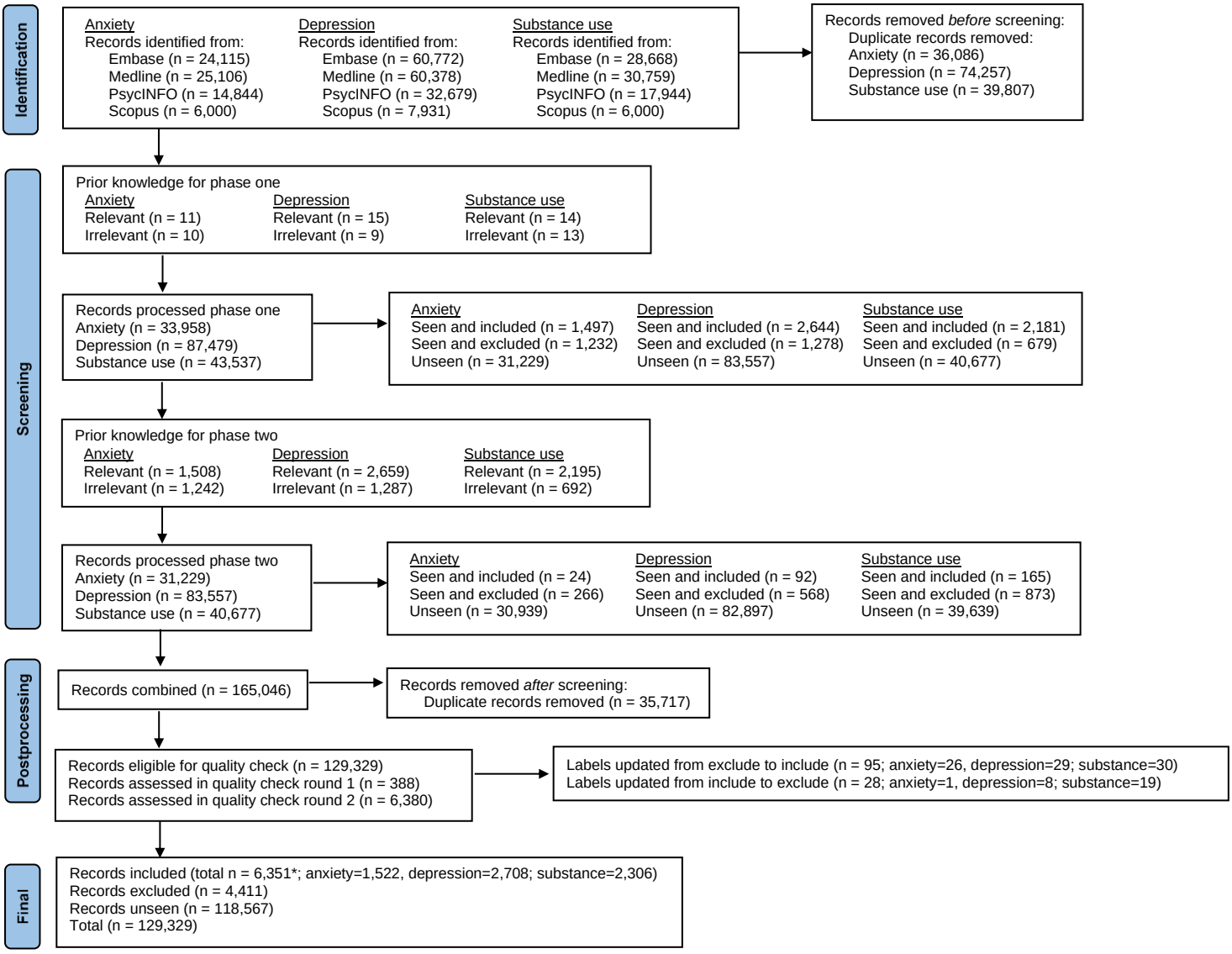

Fig 2. Modified PRISMA flowchart containing all phases conducted in this systematic review - identification, screening, postprocessing - and information about the final inclusions. The deduplicated, partly labeled (i.e., the added prior knowledge) dataset resulting from the search was used as input for the first screening phase. The output of phase one - the partly labeled dataset consisting of the priors and the labels given in this phase - was used as input for the second screening phase. The output of the second screening phase was combined, deduplicated, and the quality was checked. For more information about the phase, see the Method section.

Conclusion and Discussion

An overview of the current evidence for all published and hypothesized factors that contribute to the onset, relapse, and maintenance of anxiety-, substance use-and depressive disorders is needed to understand these CMDs, and potentially lower the great individual and societal burden. Due to the large number of studies conducted and published in these fields, an overview is time-consuming, expensive, and nearly impossible. With active learning, we were nevertheless able to create an overview of the current prospective evidence for these CMDs, resulting in a database with 6380 articles. This database will serve as a basis for international researchers interested in the current evidence of factors related to CMDs.

In the current paper, we described the process leading to the creation of this database. The current study created a partly labeled dataset regarding scientific literature on anxiety, depression, and substance abuse. However, some limitations need to be kept in mind when making use of the final dataset. Firstly, the results of simulation studies were applied to new data. For example, which model to use in the first screening phase was decided based on a simulation study performed on a different, smaller, dataset about depression (Teijema et al., 2022).

Furthermore, when to stop screening is an unresolved problem in active learning. In the current study, two stopping rules were needed - a stopping rule for phase one and phase two. In a perfect world, the stopping rule for phase one needed to be when a plateau was reached while screening (i.e., multiple consecutive irrelevant papers). In practice, the study was time-bound, making the plateau almost reached but not fully. When looking at the data, the percentage of relevant papers per session was lowering quickly, indicating that the plateau was approaching. However, the second model (i.e., a CNN model) is a powerful model, which made it possible to reach the second - final - stopping rule, regardless of the time-bound rule in the first phase. Moreover, the second stopping rule can be seen as arbitrary; there are multiple options when deciding on a stopping rule. Since the partly-labeled dataset can be used for future screening, it can be researched whether a more conservative rule would have made a significant difference regarding the final dataset.

In conclusion, this project has led to a rich database containing prospective studies that investigated preceding factors of the onset, relapse, and maintenance of anxiety, depressive, and substance use disorders. This database will receive regular updates (i.e., living review), and future research will enrich the database further. This database hence will provide a strong basis for international researchers in the field of CMDs.

References

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis [https://doi.org/10.1016/j.jocm.2018.07.002]. *Journal of Choice Modelling*, *28*, 167-182. https://doi.org/10.1016/j.jocm.2018.07.002

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry [https://doi.org/10.1136/bmjopen-2016-012545]. *BMJ Open*, *7*(2), e012545. https://doi.org/10.1136/bmjopen-2016-012545

Bramer, W. M., Giustini, D., de Jonge, G. B., Holland, L., & Bekhuis, T. (2016). De-duplication of database search results for systematic reviews in EndNote. *Journal of the Medical Library Association: JMLA*, *104*(3), 240.

Brouwer, M. (2022a). *Final data for the mega meta project* DANS. https://doi.org/10.17026/dans-z7w-9446

Brouwer, M. (2022b). *Pre-processing data for the Mega Meta Project* DANS. https://doi.org/10.17026/dans-29d-n6yg

Brouwer, M., Laura Hofstee, Jan de Boer, Felix Weijdema, Paul Lucassen, Peter M A Sloot, Karien Stronks, Julia van Weert, Reinout Wiers, Rens van de Schoot, & Bockting, C. (2021). Search Protocol for the Mega-Meta Study on Factors Contributing to Substance Use, Anxiety and Depressive Disorders. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/M5UHY

Brouwer, M. E., Williams, A. D., Kennis, M., Fu, Z., Klein, N. S., Cuijpers, P., & Bockting, C. L. (2019). Psychological theories of depressive relapse and recurrence: A systematic review and meta-analysis of prospective studies. *Clinical psychology review*, *74*, 101773.

Chen, Y., Mani, S., & Xu, H. (2012). Applying active learning to assertion classification of concepts in clinical text. *Journal of biomedical informatics*, *45*(2), 265-272. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3306548/pdf/nihms340898.pdf

Cohen, A. M., Ambert, K., & McDonagh, M. (2009). Cross-topic learning for work prioritization in systematic review creation and update [https://doi.org/10.1197/jamia.m3162]. *J Am Med Inform Assoc*, *16*(5), 690-704. https://doi.org/10.1197/jamia.M3162

Effertz, T., & Mann, K. (2013). The burden and cost of disorders of the brain in Europe with the inclusion of harmful alcohol use and nicotine addiction. *European Neuropsychopharmacology*, *23*(7), 742-748.

Fu, Z., Brouwer, M., Kennis, M., Williams, A., Cuijpers, P., & Bockting, C. (2021). Psychological factors for the onset of depression: a meta-analysis of prospective studies. *BMJ Open*, *11*(7), e050129. https://doi.org/10.1136/bmjopen-2021-050129

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, *46*(4), 1-37.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research synthesis methods*, *11*(2), 181-217.

Harrison, H., Griffin, S. J., Kuhn, I., & Usher-Smith, J. A. (2020). Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC Medical Research Methodology*, *20*(1), 7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6958795/pdf/12874_2020_Article_897.pdf

Hofstee, L., Brouwer, M., Melnikov, V., & van de Schoot, R. (2021). Screening Protocol for the Mega-Meta Study on Factors Contributing to Substance Use, Anxiety and Depressive Disorders. . *Open Science Framework*. https://doi.org/doi.org/10.17605/OSF.IO/3ZNAR

Kennis, M., Gerritsen, L., van Dalen, M., Williams, A., Cuijpers, P., & Bockting, C. (2020). Prospective biomarkers of major depressive disorder: a systematic review and meta-analysis. *Molecular Psychiatry*, *25*(2), 321-338. https://doi.org/10.1038/s41380-019-0585-z

Lefebvre, C., Manheimer, E., & Glanville, J. (2008). Searching for studies. *Cochrane handbook for systematic reviews of interventions: Cochrane book series*, 95-150.

Melnikov, V. (2021). *ASReview server setup for MegaMeta study*. In (Version v0.1.0) Zenodo. https://zenodo.org/record/5768305

Settles, B. (2012). Active Learning [https://doi.org/10.2200/S00429ED1V01Y201207AIM018]. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1-114. http://www.morganclaypool.com/doi/abs/10.2200/S00429ED1V01Y201207AIM018

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., & Silove, D. (2014). The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, *43*(2), 476-493.

Teijema, J. (2021). *ASReview CNN 17 layer model plugin* [https://doi.org/10.5281/zenodo.5084887,

Teijema, J., Hofstee, L., Brouwer, M., de Bruin, J., Ferdinands, G., de Boer, J., Siso, P. V., van den Brand, S., Bockting, C., & van de Schoot, R. (2022). Active learning-based Systematic reviewing using switching classification models: the case of the onset, maintenance, and relapse of depressive disorders.

Teijema, J., & Van de Schoot, R. (2021). *Hyperparameter-training for the Mega-Meta project*. In (Version v1.0.1) Zenodo. https://zenodo.org/record/5747050

Van de Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Ferdinands, G., Harkema, A., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Sybren, H., Tummers, L., & Oberski, D. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*, pages125–133. https://doi.org/10.1038/s42256-020-00287-7

Van de Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Ferdinands, G., Harkema, A., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Tummers, L., & Oberski, D. (2021). ASReview: Active learning for systematic reviews [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.3345592

van den Brand, S., Hofstee, L., Teijema, J., Melnikov, V., Brouwer, M., & Van de Schoot, R. (2021). *Scripts for Post-Processing Mega-Meta Screening Results*. In (Version v1.0.1) Zenodo. https://zenodo.org/record/5752358