Applying Machine Learning to Increase Efficiency and Accuracy of Meta-Analytic Review

Aaron J. Gorelik, B.S., * Mark G. Gorelik, B.S.,* Kathryn K. Ridout, M.D., Ph.D., Akua F. Nimarko, B.S., Virginia Peisch Ph.D., Shamanth R. Kuramkote B.A., Michelle Low, Tracy Pan, Simirthi Singh, Ananya Nrusimha B.A., Manpreet K. Singh, M.D. M.S**

Author Affiliations: University of California-Davis, Davis, CA (A. Gorelik); Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, FL (M. Gorelik); Department of Psychiatry, The Permanente Medical Group, Oakland, CA (Ridout); Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA (Nimarko, Peisch, M. Singh); Department of Neuroscience, Rice University, Houston, TX (Kuramkote); Department of Neuroscience, Johns Hopkins University, Baltimore, MD (Low) Department of Neuroscience, Brown University, Providence, RI (Pan); Department of Biological Sciences, University of the Pacific, Stockton, CA (S. Singh); University of California Davis School of Medicine, Davis, CA (Nrusimha)

****Corresponding author:** Manpreet K. Singh, M.D. M.S., Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, 94305; email: mksingh@stanford.edu; Phone: (650) 725-5922; Fax: (650) 724-4794

*__Equal contributions__

Summary

The rapidly burgeoning quantity and complexity of publications makes curating and synthesizing information for meta-analyses ever more challenging. Meta-analyses require manual review of abstracts for study inclusion, which is time consuming, and variation among reviewer interpretation of inclusion/exclusion criteria for selecting a paper to be included in a review can impact a study's outcome. To address these challenges in efficiency and accuracy, we propose and evaluate a machine learning approach to capture the definition of inclusion/exclusion criteria using a machine learning model to automate the selection process. We trained machine learning models on a manually reviewed dataset from a meta-analysis of resilience factors influencing psychopathology development. Then, the trained models were applied to an oncology dataset and evaluated for efficiency and accuracy against trained human reviewers. The results suggest that machine learning models can be used to automate the paper selection process and reduce the abstract review time while maintaining accuracy comparable to trained human reviewers. We propose a novel approach which uses model confidence to propose a subset of abstracts for manual review, thereby increasing the accuracy of the automated review while reducing the total number of abstracts requiring manual review. Furthermore, we delineate how leveraging these models more broadly may facilitate the sharing and synthesis of research expertise across disciplines.

Background

A meta-analysis is a statistical methodology that quantitatively synthesizes data from individual studies, allowing overall trends to be determined for a research domain.[1] Although meta-analyses are regarded as a strong systematic approach to review and synthesize evidence, they are often time- and labor-intensive, and limited by variations in researcher methodology, interpretation, and expertise.[2-4,11]This is mainly due to the need for manual human review of studies for inclusion or exclusion criteria after an initial literature review using keyword searches from databases (e.g., Medline, PsyInfo, WebofScience). During this process of reviewing hundreds to thousands of abstracts, reviewers spend 1.5 minutes per abstract on average to decide which studies to initially include in the meta-analyses.[9,10] The increasing rate of scientific publication across research domains expands the number of papers to be included in a given meta-analytic review, compounding this problem.[4,12,13] Prior work has shown that applying machine learning (ML) approaches can select abstracts for meta-analytic inclusion or exclusion with increased efficiency and similar accuracy to human reviewers.[5-8] However, these approaches have never been applied to new meta-analysis topics differing from the original training topic to see if the models can translate the same inclusion and exclusion criteria across research domains.[5–8]

Thus, we describe a ML-based protocol to improve efficiency relative to trained human reviewers by automating review of manuscript abstracts for inclusion/exclusion criteria. This protocol also leverages ML models to assist human reviewers in de novo manuscript selection for inclusion/exclusion, with the ability to modify the model sensitivity to achieve desired trade-offs between accuracy and time invested in manually reviewing abstracts. We conclude with a suggestion for the creation of meta-concept ML model repository to facilitate collaboration across research domains.

Main

To accomplish the first goal of efficient automated review of abstracts for inclusion and exclusion criteria, the ML-models were first “trained and tested” and then “evaluated” for performance. The ML models tested were curated keywords search (search), Multinomial Naïve Bayes classifier, BERT (Bidirectional Encoder Representations from Transformers), and SciBERT (BERT trained on scientific literature). BERT and SciBERT models come pretrained using a self-supervised approach from a large text corpus;[14–16] these were fine-tuned for each meta-concept (the summary of training and fine-tuning parameters and results for each ML model is summarized in Supplemental Table 1 and 2, respectively).

The “train and test” dataset contained 8202 abstracts selected for a meta-analysis examining resilience factors influencing psychopathology development in the field of psychiatry (PROSPERO protocol CRD42020172975 for details). The abstracts were evaluated for inclusion in or exclusion from four concept areas: Resilience, Biomarkers & Disease, Stressors, and Conditions. These concepts were defined using keywords and classification guidelines identified by experts in collaboration with librarians trained in systematic reviews (see Table 1 for details). Standard meta-analytic methods were used to train human reviewers to promote inter-rater agreement when reviewing for inclusion and exclusion for each concept area (see methods for training protocol).[3] These standard training methods ensured accuracy of human review and confirmed labeling needed for training four different ML models.

Table 1. Meta-concepts

Meta-Concept	Classification Guidelines	# Key Words	Example Common Terms
Resilience	The ability to adapt in the face of adversity and stress	169	Change, adapt, resilience, risk factors, adjustment
Biomarkers & Disease	Physical disorders or biological markers	282	Gene expression, oxidative stress, molecular, brain, RNA expression
Conditions	Psychiatric or neurological disorders; conditions related to brain health	563	Depression, psychopathology, anxiety, schizophrenia, internalizing
Stressors	Events or conditions in early life environment that may trigger stress	253	Stress, injury, death, trauma, SES

Table 2. Datasets used for the process. Average words and characters refers to the words and characters in the abstracts.

Data set	Source	Time frame	# Abstracts	Average words	Average characters
"Train and Test"	Embase, Web of Science, PsychInfo, CINAHL, PubMed	1976-2018	8202	211.8	1273.7
"Evaluation"	NCI	1998-2019	710	240.18	1426.8

The "evaluation" dataset contained abstracts from the field of oncology to evaluate how accurately these ML concepts transferred to a new research domain (see Table 2 for details). Tagging by a domain and meta-analysis expert (KR) was used as the "ground truth." Accuracy of the four different automation methods were compared against trained human reviewers and an untrained domain expert. The "evaluation" dataset was processed by automation, human reviewers, and the untrained domain expert without additional training. A subset of the abstracts in the "evaluation" dataset (*n*=360, 99% confidence with 5% margin of error) was compared against ground truth to ascertain relative performance. The overall process for using ML to automate abstract review is described in Figure 1.

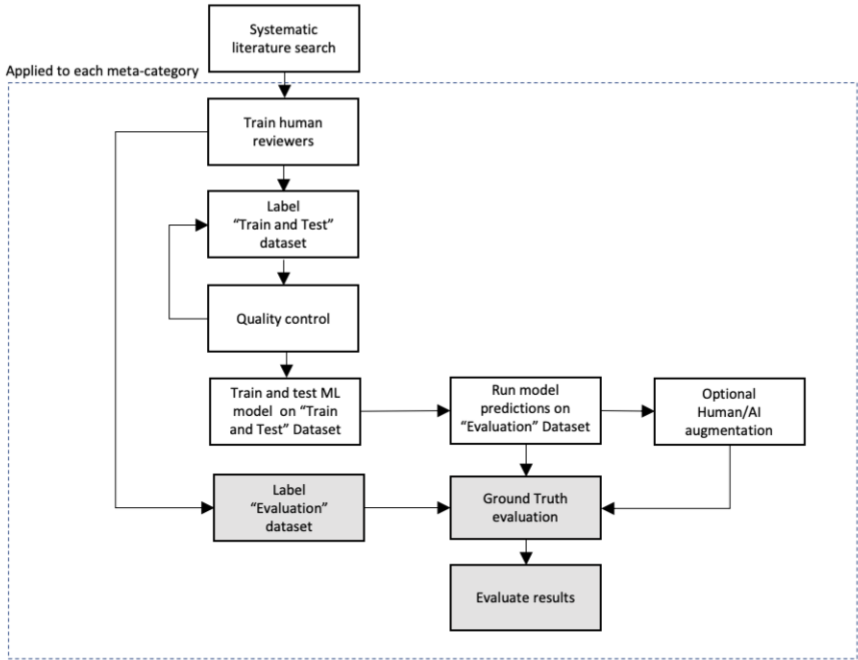

Figure 1: Process flow for individual meta-concept. "train and test" dataset was from the meta-analysis examining resilience factors influencing psychopathology development and the "evaluation" dataset was from NCI oncology dataset. The blocks in grey are only relevant to the evaluation of this methodology and would not be standard practice in real world application.

Of the four automation methods, SciBERT performed most accurately compared to human reviewers in the "evaluation" dataset, being marginally more accurate then trained human reviewers (+1% to +11%) and less accurate then the untrained expert (-3% to -11%). SciBERT had greater sensitivity for categorizing abstracts to the Biomarkers & Disease (+12.2%) and Resilience (+1.6%) meta-concepts and lower sensitivity for the Stressors (-14.3%) and Conditions (-22.5%) meta-concepts compared to the trained human reviewers (see Figure 2 for details). In addition, compared to human reviewers, SciBERT's classification performance had better precision and recall categorizing abstracts to the Conditions (F_1=+0.072), Biomarkers (F_1=+0.096), and Resilience (F_1=+0.12) meta-concepts, and slightly worse performance for the Stressors (-0.03) meta-concept (see Figure 2 for details).

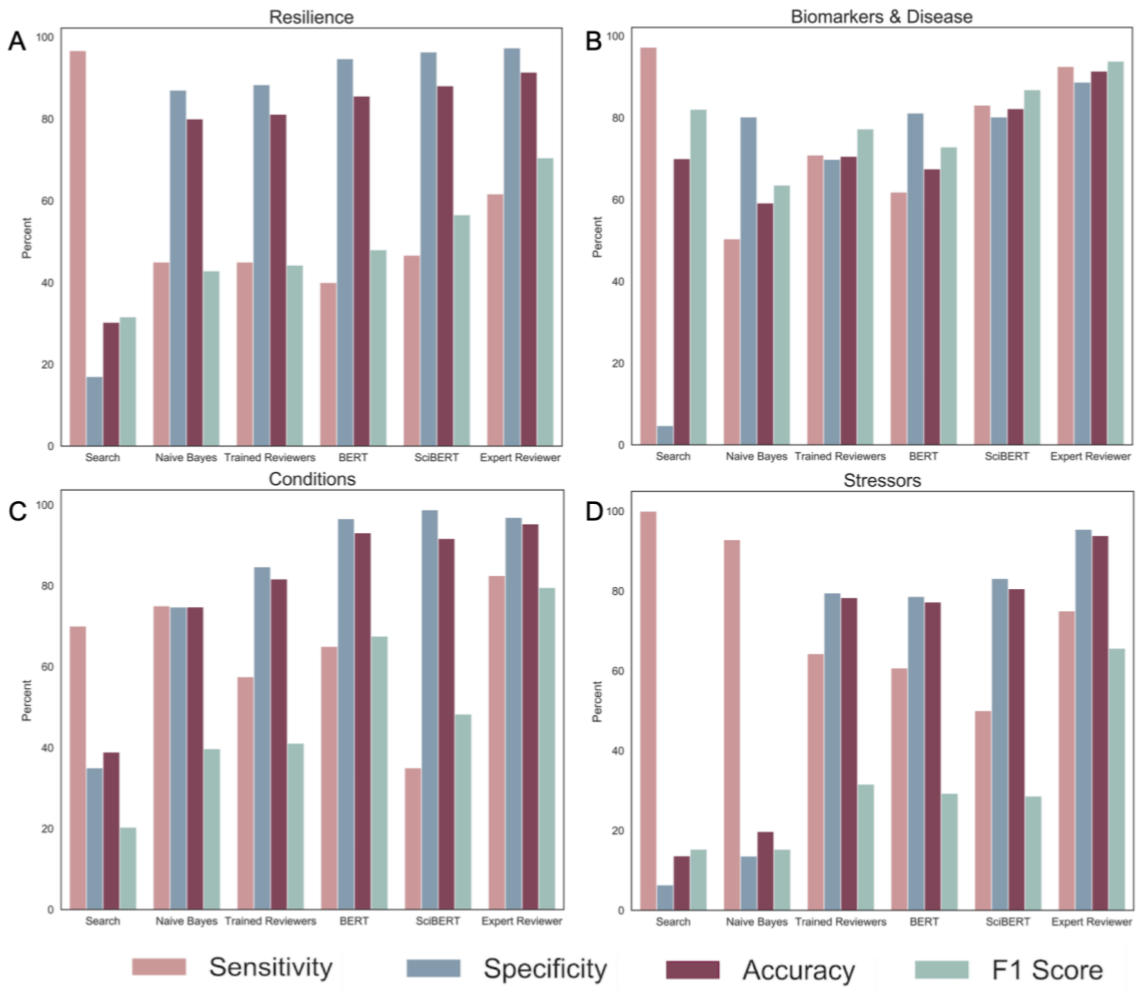

Figure 2: Paper Selection Methods Performance. Performance of all selection methods against ground truth for each meta-concept: A Resilience, B Biomarkers & Disease, C Conditions, D Stressors.

After finetuning, the SciBERT models took 0.075 seconds on average to infer inclusion/exclusion per abstract (see Table S4 for training loop and prediction times and methods for hardware details). This is compared to the trained human reviewers, who took 90 seconds on average to manually review an abstract. These results suggest that ML models trained on abstracts for specific inclusion and exclusion criteria can be translated to a new research domain while significantly reducing the time and effort required to review abstracts (Figure 2).

To accomplish the second goal of using ML models to assist human reviewers in efficient and accurate de novo abstract selection for inclusion/exclusion, we used SciBERT, which was the most similar to ground truth. SciBERT confidence scores for abstract non-inclusion (*1.0 - p(include)*) were used to recommend abstracts that needed additional human review (see Figure S3). Lower and higher SciBERT confidence ranges were tested for each concept area, and then examined for abstract classification accuracy versus number of abstracts requiring human review (please see methods for overview of process). For the low confidence range, abstracts flagged for manual review ranged from 3% of abstracts for the Conditions concept to 49% for the Stressors concept. When adding in the manual review, the accuracy and sensitivity increased by 7% and 12.5% respectively for the Conditions meta-concept, while for the Stressors meta-concept accuracy and sensitivity increased by 12% and 43%, respectively. For the high confidence range, abstracts flagged for manual review ranged from 6% for the Conditions concept to 73% for the Stressors concept. Combining the SciBERT and manual review, accuracy and sensitivity increased by 7% and 27.5% respectively for the Conditions concept, while accuracy and sensitivity increased by 15% and 46% for the Stressors concept, respectively. (see Figure 3 for all the results for both low human review and high sensitivity for each meta-concept for both augmentation types). As such, this ML approach can flag a subset of abstracts based on model confidence for manual review for inclusion/exclusion in meta-analyses, allowing human reviewers to calibrate between abstract review accuracy versus time spent manually reviewing abstracts.

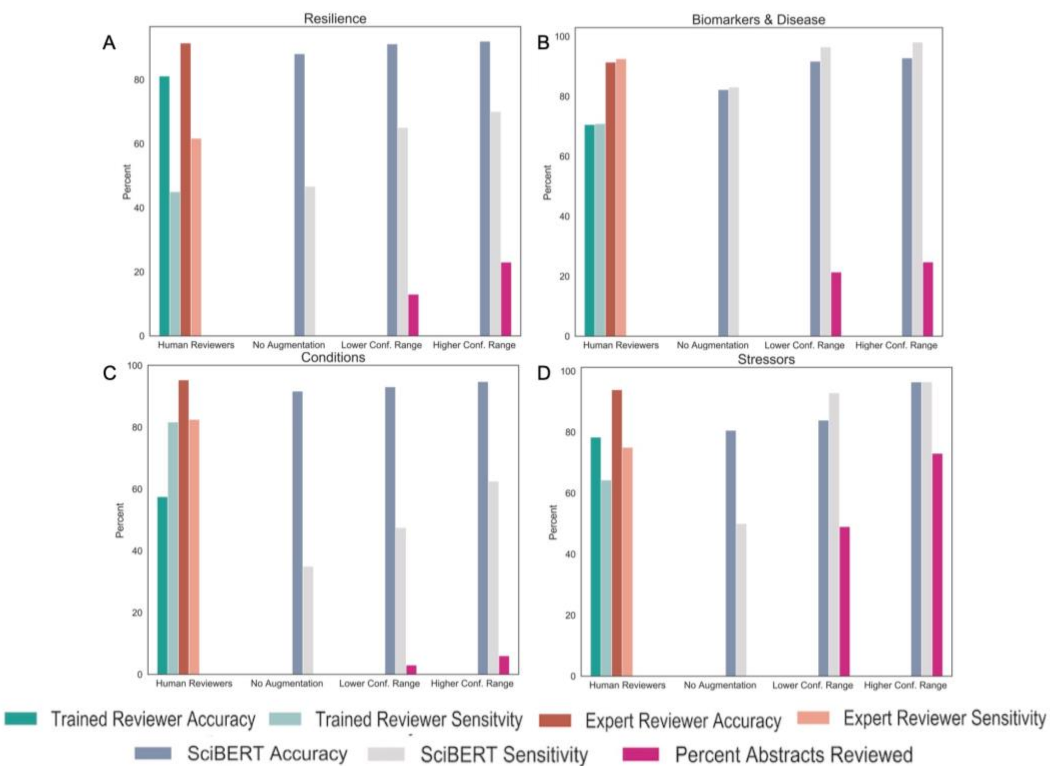

Figure 3: Human/AI augmentation performance. Shows use of artificial intelligence augmentation on volume of abstracts to review and impact on sensitivity and accuracy compared to trained human reviewers and untrained expert relative to ground truth for each meta-concept: A Resilience, B Biomarkers & Disease, C Conditions, D Stressors.

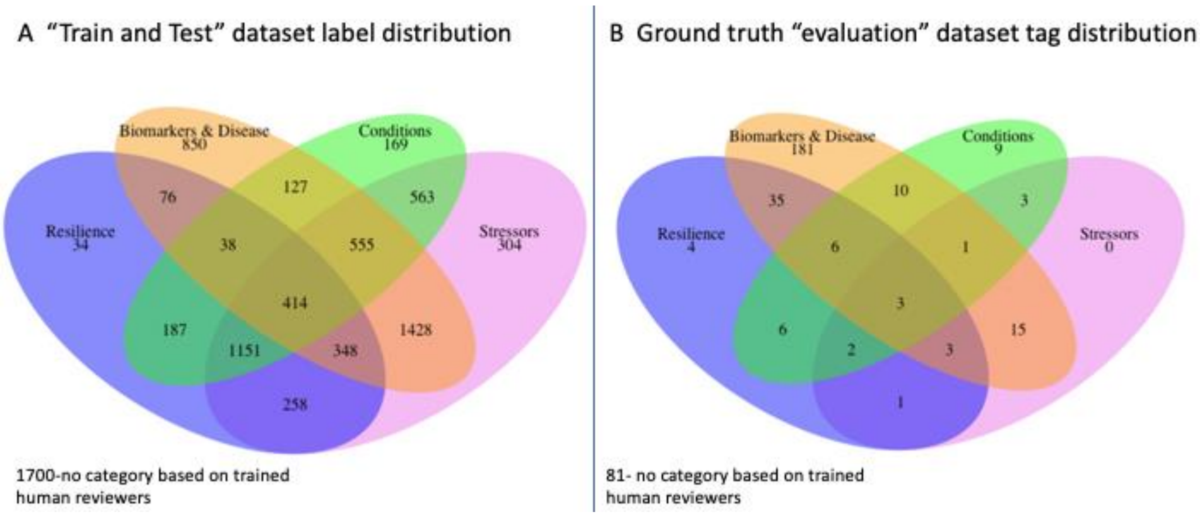

Figure 4: Meta-concept Overlaps. A. “train and test” dataset label distribution shows how n=8202 abstracts were manually tagged by trained human reviewers. B. “evaluation” dataset n=360 label distribution tagged by ground truth.

As the body of scientific knowledge grows and scientists become more specialized, evidence synthesis across disciplines is becoming increasingly challenging. Technology can be leveraged to help us organize and connect knowledge across domains. Prior work focused on similarity matching, which looks at clustering abstracts based on the similarity of words/sentences.[7,8] Similarity matching works well for systematic reviews that have narrow inclusion criteria (e.g. specific terminology) but does not support the automated transfer of meta-concept to different domains. Consequently, prior models may be less useful for either reproducibility or generalizability for future meta-analytic reviews.[7,17,18] BERT and SciBERT ML models leverage bidirectional representations to generate pretrained models drawing from a broader set of concepts. With refinement of the semantic and structural knowledge captured during the pretraining stages (transfer learning) to a substantially smaller data set, we were able to transfer ML model concepts across research domains (e.g. from psychiatry to oncology datasets).[14,15] In addition, by creating a separate ML model for each concept, we are able to use the concepts as filters that can be utilized to find papers that have any combination of these concepts (See Figure 4). This feature is particularly useful for the repurposing of these models to future meta-analyses and sets the foundation for conceptualizing ML meta-concept repositories (See Figure 4 for meta-concept interactions in the evaluation n=360 and the “train and test” datasets).

We acknowledge the limitations of our current work. For our models, ground truth was assumed to always be accurate. However, human interpretation of concepts can introduce variation into a model. With the "evaluation" dataset, we did not retrain human reviewers regarding interpretation of the four concept areas to the new research domain, which is part of standard training to ensure inter-rater reliability (this is reflected in the trained reviewer error analysis, see supplement). However, not retraining human reviewers was done intentionally to evaluate the differences in human versus ML model interpretation of the concepts across the research domains.[19–21] Further, BERT and SciBERT are computationally expensive and required truncation of text to 512 characters which could lead to potential data loss. This limitation is temporary as the most recent advancement in self-attention allows linear scaling with text size enabling our approach to apply to any size text.[22] Finally, we acknowledge that model performance can vary for different datasets. Despite these limitations, we have created an ML-based method to augment meta-analysis review and have demonstrated practical improvements in accuracy and efficiency. Further, we have shown the possibility of transferring concepts across research domains to connect and expand knowledge across scientific fields.

Future Direction

Our study provides an initial demonstration that advanced ML models can efficiently and accurately capture a meta-concept and apply the concept from a particular research field (e.g. psychiatry) to another research field (e.g. oncology). However, with these models, there is still significant work required to train reviewers, label data, and train or fine-tune a concept model. Creating a concept model repository with meta-concepts would facilitate sharing of these models and significantly reduce the efforts of researchers in creating, refining, sharing, and applying ML models across research fields, in order to easily and efficiently select papers that are relevant to meta-concepts. Additionally, this would also facilitate the creation of a common ontology that could be shared across different meta-analyses and across disciplines. For this process to be effective, standardization for creating and documenting meta-concept ML models would need to be developed. Some of the key components for this process, such as defining and evaluating meta-concept models have been described in the methods section.

Our approach demonstrates the feasibility of applying ML to meta-analytic methods to improve efficiency and accuracy within and across research fields. A repository of ML models to represent different meta-concepts could enable researchers to more broadly share and synthesize their expertise to integrate knowledge across scientific fields.

Author Contributions

AG, MG, KR, and MS conceptualized and executed the study protocol. AG, MG, and KR contributed to the analyses. AG, MG, MS, KR, VP, and AFN contributed to writing of the manuscript. AG and MG contributed to visualizations. KR, AN AFN, VP, SK, TP, ML, SS, and MS contributed to the manual tagging of abstracts.

Code availability

The code, models, and datasets are available at https://github.com/MetaAnalysisPipeline

Funding

The authors have no funding to disclose related to this work.

Conflicts of interest

Dr. Singh has received research support from Stanford's Maternal Child Health Research Institute and Department of Psychiatry, National Institute of Mental Health, National Institute on Aging, Johnson and Johnson, Allergan, Patient-Centered Outcomes Research Institute, and the Brain and Behavior Research Foundation. She is on the advisory board for Sunovion, is a consultant for Limbix, has been a consultant for Google X, and receives royalties from the American Psychiatric Association Publishing. Dr. Ridout receives support from The Permanente Medical Group's Physician Researcher Program. No other authors report any biomedical financial interests or potential conflicts of interest.

Acknowledgments

We would like to express appreciation to the members of our Pediatric Emotion And Resilience Lab (PEARL) for their contributions to this work.

Methods

Meta-Analysis Protocol and Registration

The meta-analysis examining resilience factors influencing psychopathology development protocol was registered with the International Prospective Register of Systemic Reviews (PROSPERO, CRD42020172975). The study criteria were designed using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). Portions of The Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Guidelines were also followed and adapted into PRISMA.[23,24]

"Training" meta-analysis study eligibility

Studies were included for the meta-analysis examining resilience factors influencing psychopathology development if they (1) examined the effects of early adversity in the form of abuse, neglect, socioeconomic status (SES), or other adverse exposures on human subjects occurring prenatally up to age 18; and (2) provided adequate description of adversity assessments. Studies using indirect proxies of early adversity, such as parental education alone, were excluded. Prospective, observational, and retrospective studies were considered.

Meta-Analysis Information Sources and Search Strategy

For the meta-analysis examining resilience factors influencing psychopathology development, a comprehensive electronic search conducted in September 2018 identified English language studies indexed in PubMed/Medline, PsycINFO, CINAHL, and Web of Science; no publication date limitation was set. The search was performed by investigators with topic clinical and research experience (KR) in consultation with a librarian trained in systematic reviews. The search strategy included terms and combinations to identify early life stress, biomarkers and resilience (See Supplementary Table 3 for search criteria). Primary study and review article references were searched; studies were appraised for inclusion or exclusion using an a priori criteria as described under study eligibility.

Datasets

Two different abstract datasets were utilized: a "train and test" data set for training the ML models (obtained from the meta-analysis examining resilience factors influencing psychopathology development) and an "evaluation" data set (oncology research field) to evaluate the ML models' performance (see Table 2 for dataset details). For the "evaluation" (oncology) dataset, we used the NCI cancer control publications from 1995-2019 where author type included was "Both NCI & Extramural Researchers."

Concept inclusion and exclusion criteria

A definition was created for each of the four concepts based on key terms developed for the "train and test" meta-analysis (see Table 1) and expanded by content experts (MKS and KR). For the "training" dataset, human reviewers were trained regarding inclusion and exclusion criteria for each concept as found in the reviewer training and dataset labeling section below. For the "evaluation" dataset, ground truth, the human reviewers, and the untrained domain expert were asked use their "judgment and intuition" regarding the four concepts in addition to searching the abstracts for the presence of key terms. Table 1 describes the criteria for each meta-concept (see supplement table 4a-4d for list of key words).

Reviewer Training and dataset labeling

The manual review team (SK, ML, AFN, AN, TP, VP, and SS) consisted of members from varying degrees of domain expertise, including undergraduate, graduate, postdoctoral, and faculty level coders.[25] After a group training regarding the goals of the study, the study inclusion/exclusion criteria, and concepts, the team tagged 150 identical abstracts obtained from the "train and test" data set, and then compared their results to ground truth. Any discrepancies in labeling were resolved in a joint training session. This training resulted in >90% inter-rater reliability. In addition, the reviewers went through an extensive training protocol including a quality control (QC) methodology to ensure consistent results (see QC section). Reviewers labeled both the "train and test" and "evaluation" abstract datasets. A research field expert and expert in meta-analysis methodology (KR) was used as the ground truth in validating the manual tagging results from the "train and test" dataset for abstracts where there was a difference in labeling between trained reviewer and ground truth, and as a final arbiter for any ambiguity. MKS, our untrained domain expert reviewer, did not undergo the aforementioned reviewer training or QC process.

The QC process and reviewer training was applied only to the "train and test" data sets. This was to allow for evaluation and comparison of meta-concept translation by humans compared to the ML models.

Quality Control (QC)

The unlabeled abstracts in the "train and test" dataset were divided equally among the trained reviewers. Once a human reviewer had labeled their portion (~1,350 abstracts), 150 random abstracts were checked against ground truth (99% confidence with 5% margin of error). If the percent agreement rate with ground truth was below 90%, that reviewer's portion was relabeled by two new reviewers after which 150 samples (99% confidence with 5% margin of error) were re-checked against ground truth. Only one reviewer's result was below 90% agreement and was therefore relabeled and re-validated. In addition to QC, any reviewer could discuss abstracts in the "train and test" data set to the ground truth to resolve ambiguity. However, in the "evaluation" dataset (not part of the standard meta-analysis protocol), these QC protocols were not utilized to allow comparison of meta-concept application between human reviewers and the ML model.

Automated Filters

Four different automated approaches for selecting papers were tested and validated: one keyword search and three machine learning based models (Multinomial Naive Bayes, BERT, SciBERT).[14,15,26]

Multinomial Naïve Bayes is a commonly used text classification approach that has been shown to perform well in applications such as spam filtering.[27] We used the scikitlearn implementation of Naïve Bayes due to its widespread use. The Naïve Bayes model was configured to get best performance for each meta-concept. The model was trained on a term frequency (TF) normalized matrix of word counts from the training set for Stressors and Conditions meta-concepts. For Resilience and Biomarkers & Disease meta-concepts the model was trained on term frequency–inverse document frequency (TF-IDF) normalized matrix of word counts from the training set (for details of other parameters see Table S1 in supplement). The model was optimized using a grid search.

BERT and SciBERT have proven to be extremely powerful for picking up context dependent patterns that may be missed by more traditional approaches, such as Naïve Bayes. BERT and SciBERT leverage bidirectional representations to generate pretrained models. BERT and SciBERT both came pre-trained on vocabularies of approximately 30,000 unique words and subwords. The main difference between the two is that BERT is trained on general purpose text corpus (book corpus and Wikipedia), whereas SciBERT is trained on scientific literature.[14] The texts used to pre-train BERT and SciBERT share approximately 42% of the unique words. All four meta-concepts used the same parameters (see table S2 in supplement). Pre-trained BERT and SciBERT models were downloaded from huggingface.[28]

All ML based models were trained and measured on the same training test split of the "train and test" dataset implemented in sklearn using a seed value of 42 to ensure reproducible results. The split was 90% training and 10% test data. A smaller test portion was chosen since the real validation was done against a separate "evaluation" dataset. The Multinomial Naïve Based model was trained from scratch, whereas BERT and SciBERT were fine-tuned on top of pretrained models based on different text corpuses. For fine-tuning, the abstract texts were truncated at 512 words (see supplement Figure 1 and 2 for details).

For keyword search testing, the keyword lists were constructed by domain experts for the keyword search testing model. All keywords were at least 3 characters and abstracts were included if they contained at least one keyword. Special characters, such as asterisks, were removed from the search terms in order to mitigate inconsistent usage.

BERT and SciBERT were fine-tuned and used on the Google Colab Pro environment using a Tesla P100-PCIE-16GB. All other code was executed on a MacBook Pro retina with 2.9 GHz i9 CPU.

Combining Multiple Meta-Concepts

By creating a separate model for each concept, we were able to evaluate how often these concepts were studied together or individually. In effect, the concept became a search criterion that could be combined in a logical expression for searching for abstracts that span different categories (e.g. includes concept Resilience but not Biomarkers & Disease or Conditions). Thus, this enables researchers to use one dataset to confirm multiple possible combinations of concepts.

Result Analysis

All abstracts in the evaluation data set were labeled using automated filtering methods. Accuracy, recall, and F_1 score for all automated filtering approaches were calculated using sklearn implementation. Vectors were extracted from both the fine-tuned BERT and SciBERT models and dimensionality reduction to demonstrate generalizability using transfer learning approach was performed by UMAP (see supplement figure 4).[29]

All abstracts in the evaluation data set were labeled by a group of trained human reviewers. In addition, 360 random abstracts were selected from the "evaluation" dataset to achieve 99% confidence with 5% margin of error accuracy. The abstracts were labeled by an untrained expert and ground truth.

The results for 360 random abstracts were compared across all the different filtering/labeling methods for each meta-concept topic against ground-truth: untrained expert, trained human reviewers, key word search, and ML models: Naive Bayes, BERT and SciBERT. As an extra cautionary step, discrepancy in the 360 labels between ground truth and untrained domain expert were reviewed to reduce potential for ground truth errors.

We used the following metrics to evaluate performance of each method:

Legend: *TP* – True Positives, *FP* – False Positives, *TN* – True Negatives, *FN* – False Negatives

$Accuracy = (TP+TN) / (TN + TP + FN + FP)$

$Specificity = TN / (TN + FN)$

$Sensitivity = TP /(TP + FP)$

$F_1 = (2*TP) / (2*TP + FN + FP)$

AI/Human Augmentation

To enable the researchers to achieve a higher level of accuracy and sensitivity than a pure automated abstract selection process, the system generated a recommended list of abstracts to review. We tested low and high confidence ranges (defined below). This methodology was applied to the SciBERT model only, although the same methodology could be extended to BERT and Naïve Bayesian Classifier.

The ML model chose abstracts for manual human review within a range of negative class inference confidence values (*1-p(inclusion)*). For the lower confidence range, we used confidence values between 0.5 and 0.9, and for the higher confidence range we used values between 0.5 and 0.95. The abstracts with confidence below 0.5 were treated as true positives and values above 0.9 or 0.95 were treated as true negatives. For deep neural networks, the confidence values are typically clustered at the higher end of the confidence score (*p(inclusion)*).[30] We chose to focus on sensitivity as the key evaluation in addition to accuracy. This also aligned well with the meta-analysis process where false positives would likely be weeded out at later stages of a meta-analysis.

We used ground truth for the evaluation dataset with the assumption that a human reviewer would always label the abstracts accurately for the augmentation step. We chose not to incorporate the accuracy of ground truth because the error for human reviewers depends on quality control applied and in a meta-analysis, which would be considered ground truth.

References

1. GLASS, G. V. Primary, Secondary, and Meta-Analysis of Research. *Educ. Res.* **5**, 3–8 (1976).
2. Burns, P. B., Rohrich, R. J. & Chung, K. C. The levels of evidence and their role in evidence-based medicine. *Plast. Reconstr. Surg.* **128**, 305–310 (2011).
3. Berlin, J. A. & Golub, R. M. Meta-analysis as evidence: building a better pyramid. *JAMA* **312**, 603–605 (2014).
4. Bown, M. J. & Sutton, A. J. Quality control in systematic reviews and meta-analyses. *Eur. J. Vasc. Endovasc. Surg. Off. J. Eur. Soc. Vasc. Surg.* **40**, 669–677 (2010).
5. Lerner, I., Créquit, P., Ravaud, P. & Atal, I. Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses. *J. Clin. Epidemiol.* **108**, 86–94 (2019).
6. Bao, Y. *et al.* Using Machine Learning and Natural Language Processing to Review and Classify the Medical Literature on Cancer Susceptibility Genes. *JCO Clin. Cancer Inform.* **3**, 1–9 (2019).
7. Xiong, Z. *et al.* A Machine Learning Aided Systematic Review and Meta-Analysis of the Relative Risk of Atrial Fibrillation in Patients With Diabetes Mellitus. *Front. Physiol.* **9**, 835 (2018).
8. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
9. Michelson, M. & Reuter, K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemp. Clin. Trials Commun.* **16**, 100443 (2019).
10. Greco, T., Zangrillo, A., Biondi-Zoccai, G. & Landoni, G. Meta-analysis: pitfalls and hints. *Heart Lung Vessels* **5**, 219–225 (2013).
11. Nakagawa, S., Noble, D. W. A., Senior, A. M. & Lagisz, M. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC Biol.* **15**, 18 (2017).
12. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).
13. Elliott, J. H. *et al.* Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Med.* **11**, e1001603 (2014).
14. Beltagy, I., Lo, K. & Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *ArXiv190310676 Cs* (2019).
15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs* (2019).
16. Vaswani, A. *et al.* Attention Is All You Need. *ArXiv170603762 Cs* (2017).
17. Norman, C. R., Leeflang, M. M. G., Porcher, R. & Névéol, A. Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *Syst. Rev.* **8**, 243 (2019).
18. Bannach-Brown, A. *et al.* Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Syst. Rev.* **8**, 23 (2019).
19. Frénay, B. & Kabán, A. A Comprehensive Introduction to Label Noise. *Comput. Intell.* 10 (2014).
20. Delgado-Rodriguez, M. Bias. *J. Epidemiol. Community Health* **58**, 635–641 (2004).
21. Song, H., Kim, M., Park, D. & Lee, J.-G. Learning from Noisy Labels with Deep Neural Networks: A Survey. *ArXiv200708199 Cs Stat* (2020).
22. Wang, S., Li, B. Z., Khabsa, M., Fang, H. & Ma, H. Linformer: Self-Attention with Linear Complexity. *ArXiv200604768 Cs Stat* (2020
23. Liberati, A. *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann. Intern. Med.* **151**, W65-94 (2009).
24. Stroup, D. F. *et al.* Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* **283**, 2008–2012 (2000).
25. van Wely, M. The good, the bad and the ugly: meta-analyses. *Hum. Reprod.* **29**, 1622–1626 (2014).
26. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
27. Kanaris, I., Kanaris, K., Houvardas, I. & Stamatatos, E. WORDS VS. CHARACTER N-GRAMS FOR ANTI-SPAM FILTERING. 20.
28. Wolf, T. *et al.* HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv191003771 Cs* (2020).
29. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2018).
30. Nguyen, A., Yosinski, J. & Clune, J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *ArXiv14121897 Cs* (2015).

Supplementary Methods & Extended Figures

Multinomial Naïve Bayes:

For any configuration values for scikitlearn package not listed in Table S1, the default values that came with the package were used.[1]

Table S1. Final Naïve Bayes Model Parameters

Feature	Resilience Model	Biomarker & Diseases Model	Stressor Model	Conditions Model
N-gram Range	(1,3)	(1,2)	(2,2)	(2,2)
TF-IDF or TF	TF-IDF	TF-IDF	TF	TF
alpha	.0001	.001	.001	.0001

BERT & SciBERT:

The parameters used to fine-tune the BERT and SciBERT models were the default recommendations based on the authors of BERT (Table S2).[2,3] For any parameters not listed in Table S2, the default values were used that came with the huggingface package.[4]

Table S2. Final BERT and SciBERT model parameters

Parameters	BERT	SciBERT
Model	bert-base-uncased	allenai/scibert_scivocab_uncased
Learning Rate	$2e^{-5}$	$2e^{-5}$
Adam Epsilon	$1e^{-4}$	$1e^{-4}$
Batch size	8	4
Epochs	4	4
Text Truncation Cutoff	512	512
Seed	42	42

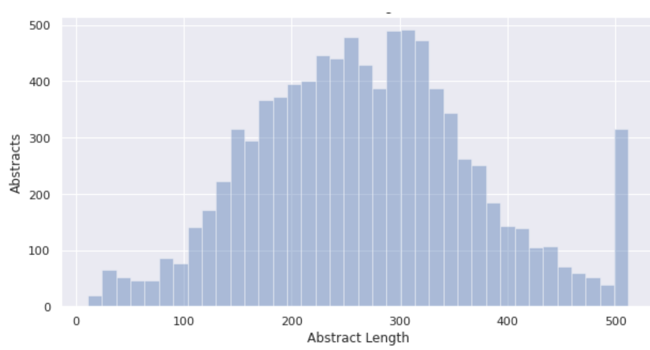

Fig S1. "train and test" Dataset Abstract Character Count Distribution (Post Truncation). Abstracts used for BERT and SciBERT were truncated at a maximum length of 512. Most abstracts were 200-400 characters long in the train/test dataset. Roughly 300 abstracts needed to be truncated.

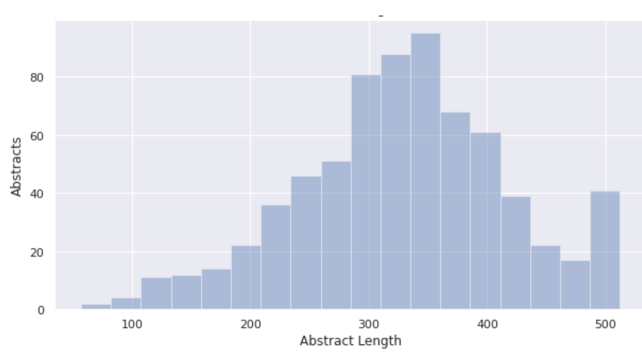

Fig S2. "evaluation" Dataset Abstract Character Count Distribution (Post Truncation). Abstracts used for BERT and SciBERT were truncated at maximum length of 512 characters. Majority of abstracts were 300-400 characters long in the evaluation dataset. Only 40 abstracts needed to be truncated.

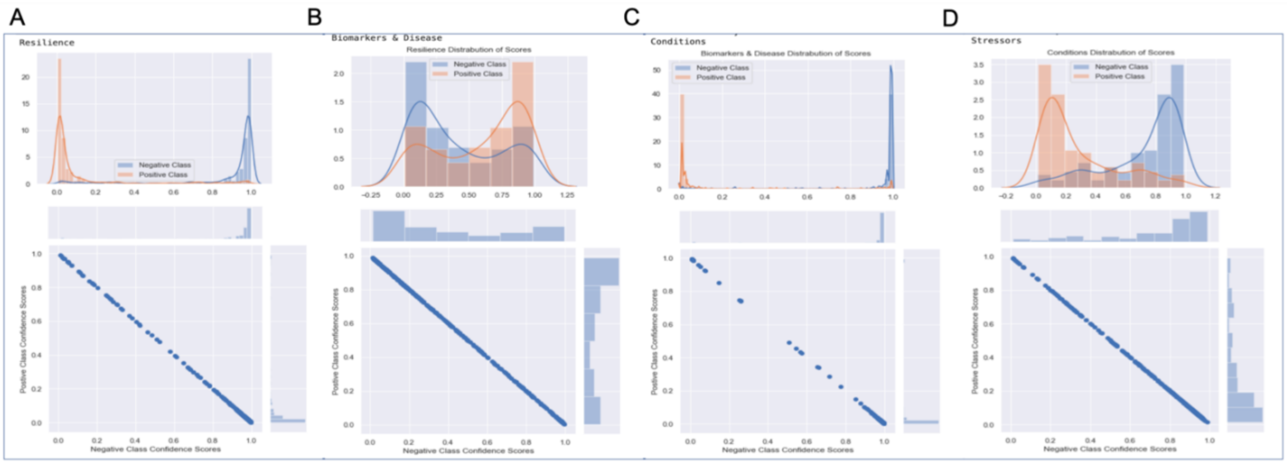

Fig S3. Confidence value distributions of predictions by SciBERT from the "evaluation" dataset. Concepts the model performed well on during the training, such as the Resilience and Conditions categories, had more extreme distributions of scores. Lower scoring categories, such as Biomarkers & Diseases and Stressors, had lower confidence predictions. This suggests that during Artificial Intelligence (AI)/Human augmentation, these categories would need more manual intervention.

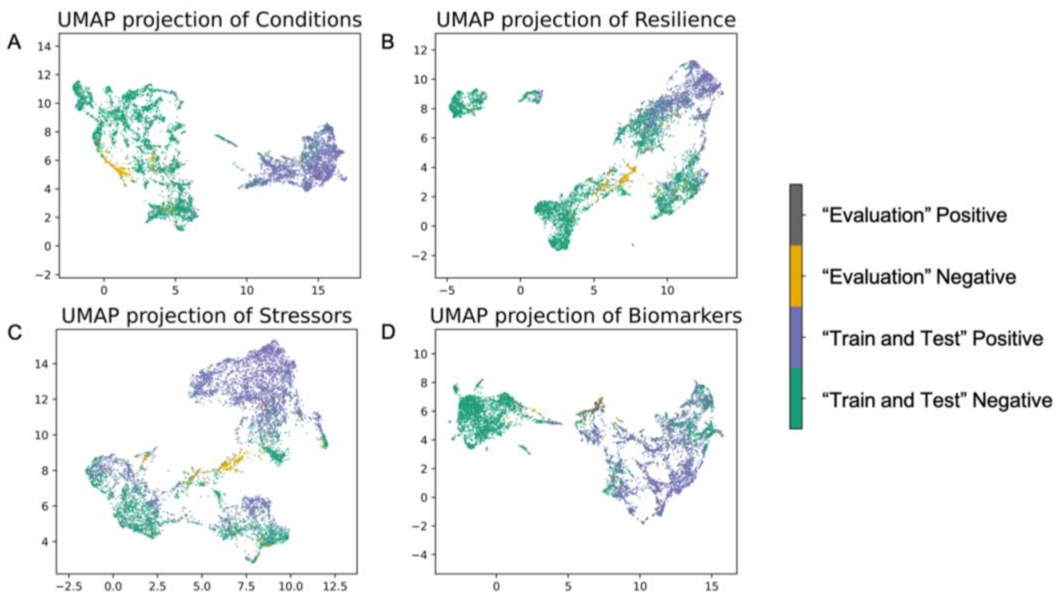

Fig S4. UMAP of embeddings from SciBERT of the "Train and Test" Dataset and subset (n=360) of the "Evaluation" Dataset. There was consistent clustering of positive and negative classes. The evaluation data set, overall, cluster with their respective test/train classes suggesting the ability of the model to incorporate novel corpuses of relevant academic literature.

Table S4 Training loop times and prediction times for SciBERT. Average prediction time was calculated by measuring how long it took to predict the entire evaluation set and dividing by the length of the evaluation set.

Category	"Train and Test" dataset training time (seconds)	"Evaluation" dataset prediction time (seconds)	Average time for a single prediction (seconds)
Resilience	2128.44	13.67	0.0189
Biomarkers & Disease	2103.46	13.52	0.0187
Conditions	2103.83	13.52	0.0187
Stressors	2107.16	13.51	0.0187

Trained Reviewers:

The consistency and performance of the trained reviewers on meta-concepts varied on sensitivity, specificity, accuracy, and F_1 scores (Fig S5). Reviewer performance did not correlate with their education, age, or expertise on the scientific domain. Individual reviewer performance was also not consistent across meta-concepts. Even with extensive training the reviewers' performance was neither 100% accurate nor consistent across meta-concepts compared to ground truth. ML models can perform on par with or better than trained human reviewers. By combining ML and human reviewer efforts as outlined by our human augmentation protocol, we can improve the precision of the abstract review and inclusion process to achieve close to untrained domain expert accuracy and precision.

It is important note that the reviewers were trained and quality-controlled on the "train and test" (psychiatry) dataset and were not provided any additional guidance for transferring the meta-concepts learned in the "train and test" test dataset to the "evaluation" (oncology) dataset.

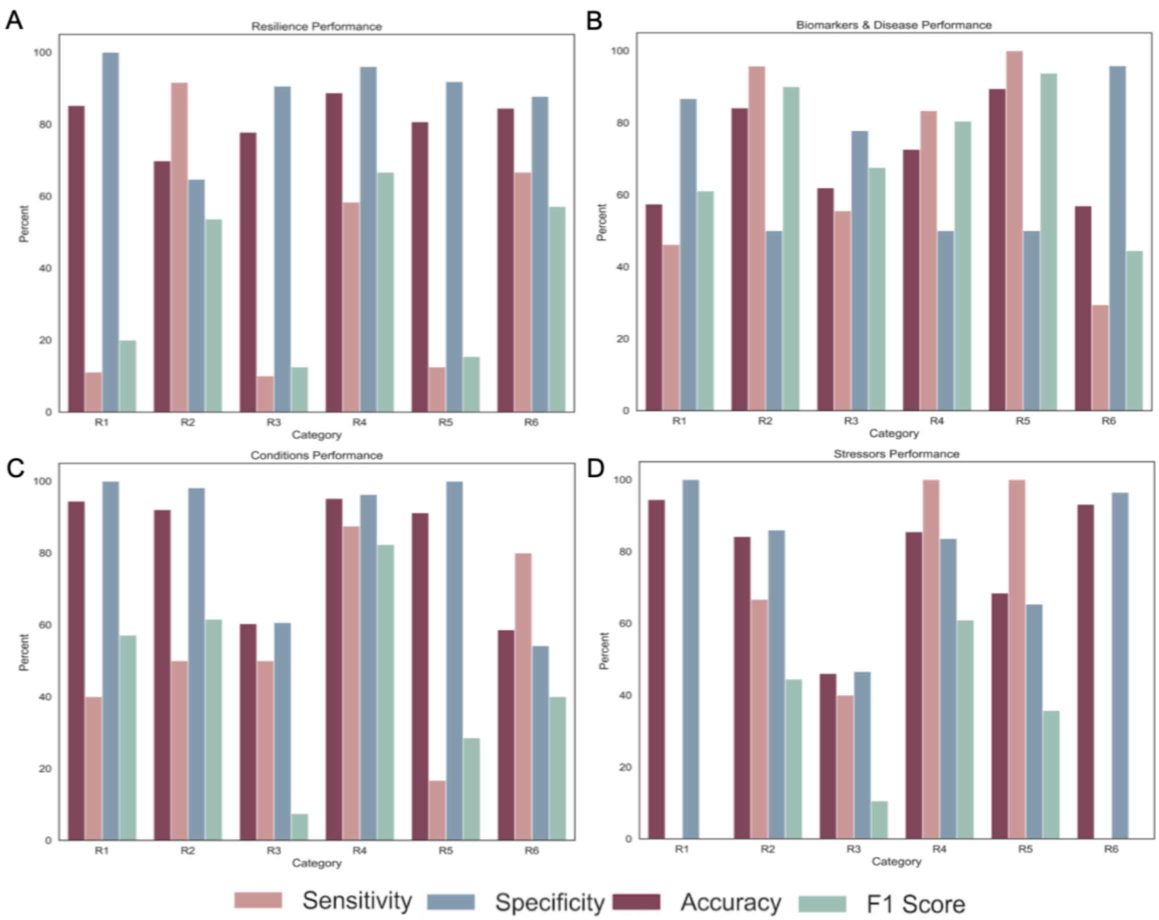

Fig S5. Performance of trained reviewers on the randomly sampled (*n*=360) "Evaluation" dataset. The reviewers' performance were not consistent across education level, age, expertise or concepts.

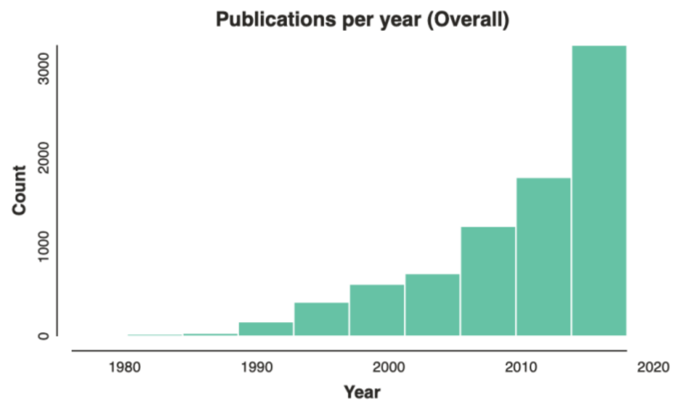

Fig S6. Publications over time in "train and test." This is based on expert keywords extracted for the meta-analysis examining resilience factors influencing psychopathology development.

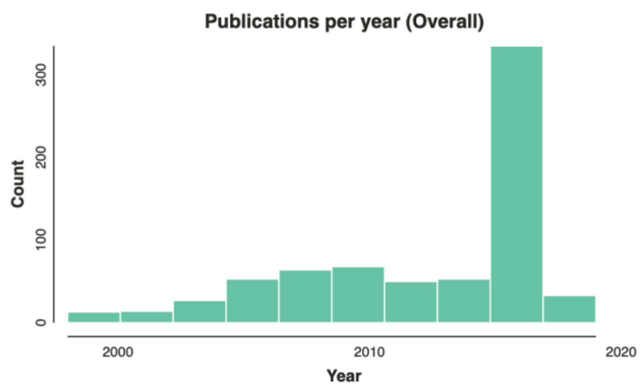

Fig S7. Publications over time in "evaluation" from curated NCI Dataset *n*=720.

Supplemental Table 3. Full search strings

Search strategy	Search was performed from the earliest available date up to July 2016. The following (key)words and MeSH terms, including combinations, were used: Child Abuse, Maternal Deprivation, Paternal Deprivation, Family Conflict, Maternal-Fetal Relations, Battered Child Syndrome, Life Change Events, Adult Survivors of Child Adverse Events, Child, Orphaned, Child, Foster, Homeless Youth, Neglect, Maltreatment, Abandonment, Parental loss, Maternal separation, Maternal depression, Adverse childhood experience, Early adversity, Early life stress, Resilience, Psychological, Protective Factors, Risk Factors, Resilience, Grit, coping skill coping method, Hardiness, Risk factors, Biomarkers, Pituitary-Adrenal System, Sympathetic nervous system, Epigenesis, Genetic Allostasis Oxidative stress, DNA Damage, Epigenomics, Hydrocortisone/blood, Stress, Physiological/blood, Mental Disorders, Psychopathology, Cardiovascular Diseases/etiology, Digestive System Diseases/etiology, Endocrine System Diseases/etiology, Immune System Diseases/etiology, Musculoskeletal Diseases/etiology, Nervous System Diseases/etiology, Respiratory Tract Diseases/etiology, Chronic Disease/etiology, Nutritional and Metabolic Diseases/etiology, Disease, Chronic illness, Diabetes, Hypertension, Heart disease, Asthma, Depression, Mental illness, mood disorder, cyclothymia, Bipolar, Schizophrenia, seasonal affective disorder, PTSD, post-traumatic stress disorder, Dysthymia, Anxiety, premenstrual dysphoric disorder, PMDD, Mood disorder			
Database	PubMed/Medline	PsycINFO & CINAHL	Web of Science	Embase
Filter/limit	-	-	"article"	
Number of articles found	3,944	4,846	1,135+3,283	1096
Unique articles within database	-	-	-	-

Supplemental Table 4a. Full key terms for search method: Resilience

Key Terms	Acceptance, Achievement, Achieve goals, Achieve my goals, Act on a hunch, Act on hunch, Active coping, Adapt, Adaptable, Adaptability, Adjust, Attain goals, Best effort, Bounce back, Challenge, Change, Changes, Close and secure relationship, Close relationship, Cognitive control of emotion, Cognitive flexibility, Cognitive reappraisal, Commitment ,Confidence, Connection, Control, Cope, Coping, Coping mechanism, Coping methods, Coping skill, Coping with stress, Core belief, Deal with everything, Difficult times, Does not take me long to recover from a stressful event, Does not take me long to recover,

	Efficac*, Efficacy, Emotional intelligence, Empowered, Empowerment, Endurance, Even when hopeless do not give up, Do not give up, Failure, Family coherence, Fears, Feel in control, Flex*, Flexibility, Fortitude, Grit, Handle unpleasant, Handle unpleasant feeling, Hard time making it, Hard time making it through stress, Hardiness, Hardy, Hardi*,Hope*, Integration, Internal locus of control, Intervention, Know where to find help, Like challenge, Like challenges, Make unpopular, Make unpopular decisions, Mastery, Mental muscle, Mental power, Morals, Moral integrity ,Most things happen for a reason, Not easily discouraged, Not discouraged, Openness, Optimism ,Organizing ,Past success, Performance, Perspective, Physical endurance, Physical hardiness, Physical wellbeing, Physical well-being ,Physical well-being, Planning, Plasticity, Positive attitude, Positive emotions, Positive statement, Positive statements, Prefer to take lead, Prefer to take the lead, Prevention, Productivity, Protect, Protective, Protective Factors, Psychological capitol, Psychological resiliency, Takes lead, Takes the lead, Recover, Recovery, Religion, Resilience, Resilience Psychological, Resilience Psychological Resiliency, Resiliency, Resist , Resource*, Resourceful*, Risk Factors, Role model, Role models, Safety net, Secure relationship, See humor, Self-advocacy, Self awareness, Self-aware, Self-awareness, Self-differentiation, Self-efficacy, Self efficacy, Self esteem, self-esteem, Set-backs, Set backs, Signature strengths, Social competence, Social support, Snap back, Snap-back, Spiritual belief, Spirituality, Stay focused, Strategy, Strength, Stress inoculation, Strong person, Strong sense of purpose in life, Strong sense of purpose, Structured environment, Success, Support, Supportive relation*,Survivor mission, Take pride, Take pride in achievements, Takes lead, Takes the lead, Therapy, Think of myself as a strong person, Times of stress know where to find help, Tolerance, Traumatic, Treatment, Trust, Under pressure stay focused, Value, Wellbeing, Well being, Well-being, Wellness, Work to attain goals

Supplemental Table 4b. Full key terms for search method: Biomarkers & Disease

Key Terms	5-HT, ACC, Acetyl L carnitine, Acetyl-L-carnitine, Adrenal, Adrenaline, Adrenal*, Adreno, Adreno*, Allostasis, Allostatic load, Asthma, alpha amylases, alpha-Amylases, Amygdala, Anterior cingulate cortex, Astma, Autonomic nervous system, Basal ganglia, BDNF, Biomarker, Biomarkers, Blood pressure, Blood, BMI, Body mass index, Brain, Brain derived neurotrophic factor, Brain-derived neurotrophic factor, Bronchial asthma, CA 1, Cardiovascular, Cardiovascular disease, Cardiovascular diseases, Cardiovascular disease etiology, Cardiovascular diseases/etiology, Cardiovascular etiology, Caudate nucleus, CDH13, cerebrovascular, Chemical, Chronic , Chronic disease, Chronic disease etiology, Chronic disease/etiology, Chronic etiology, Chronic illness, Cortico, Cortisol, Cortico*, CSIF, CVD, Cytokin, Cytokine synthesis inhibitory factor, Cytokines, Cytokin*, C elegans, C-reactive protein, C. elegans, $C_6H_{12}O_6$, Dehydroepiandrosterone sulfate, Dehydroepiandrosterone-sulfate, DHEA-S, Diabetes, Digestive, Digestive system , Digestive system disease, Digestive system disease etiology, Digestive system diseases, Digestive System diseases/etiology, Digestive system etiology, Disease, DLPFC, DNA damage, DNA methylation, DNA methyltransferase, Dopamine, Dorsolateral prefrontal cortex, Dti, Dysregulation, Endocrine, Endocrine system, Endocrine system disease etiology, Endocrine system diseases/etiology, Endocrine system etiology, Enzyme, Epigenetic, Epigenesis, "Epigenesis, Genetic," Epigenomics, Epinephrine, Etiology, FGF2, Fibrinogen, FKBP Prolyl Isomerase 5, FKBP5, FK506 binding protein 5, FMRI, Freesurfer, Frontal cortex, Gene, Gene Expression, Genetic, Gland, Globus pallidus, Glucocorticoid, Glucose, Glycosylated hemoglobin , Glycated hemoglobin, HbA1C, HBP, Heart disease, Heart rate, Heart rate variability, Hemoglobin A1C, HDL, High blood pressure, High density lipoprotein, High-density lipoprotein, Hippocampus, Homeostasis, Hormone, Hormones, HPA axis, Human, Human cytokine synthesis inhibitory factor, Humans, Hydrocortisone, Hydrocortisone/blood, Hydrocortisone/blood*, Hypertension, Hypothalam, Hypothalamic Pituitary Adrenal axis, Hypothalamic-Pituitary-Adrenal axis, Hypothalam*, Illness, Illnesses, IL1F2, IL-1b, IL-1beta, IL-1 beta, IL-10, IL-6, Immune, Immune system, Immune system disease, Immune system diseases, Immune system diseases etiology , Immune system disease/etiology, Inflammation, Inflammation mediators, Interleukin 1 beta, Interleukin-1 beta, Interleukin 10, Interleukin-6, Interleukin-10, Insula, Insulin, Insulin resistance, Interleukin 6 , Interleukin-6, LAC, LDL, Lead, Lead intoxication, L acetylcarnitine, L-acetylcarnitine, Ligand, Limbic, Locus coeruleus, Low density lipoprotein, Low-density lipoprotein, Macaque, Magnetic Resonance Imaging, Messenger RNA, Metabolic , Metabolic disease, Metabolic disease etiology, Metabolic diseases, Mice, Microbiome, MicroRNA, MiRNA, Mitochondria, Mitochondrial DNA copy number, Molecular, Monkey, Mouse, MRI, mRNA, mtDNAcn, Murine, Musculoskeletal , Musculoskeletal disease, Musculoskeletal diseases, Musculoskeletal disease etiology, Musculoskeletal diseases/etiology, Neural correlates, Neurobehavioral, Neurobiolog, Neurobiology, Neurobiolog*, Neurochemical, neurocircuitry, Neurophysio, Neurophysiological, Neuroscience, Nervous system, Nervous system disease, Nervous system diseases, Nervous system disease etiology, Nervous system diseases/etiology, NRG1, NR3C1, Nuclear Receptor Subfamily 3 Group C Member 1, Nucleus accumbens, Nutrition and metabolic disease etiology, Nutrition and metabolic diseases, Nutritional , Nutritional and metabolic, Nutritional and metabolic diseases/etiology , Nutritional disease, Nutritional diseases, Nutritional disease etiology, Oxidative stress, Oxytocin, Pituitary adrenal, Pituitary adrenal System , Pituitary-adrenal, Pituitary-adrenal System, Plasma , Plasticity, PND40, Prefrontal cortex, Pr interval, Proinflammatory cytokines, Pro-inflammatory cytokines, Putamen, Rat, Reactive airway disease, Receptor, Respiratory, Respiratory tract, Respiratory tract disease, Respiratory tract diseases, Respiratory tract diseases/etiology, Respiratory tract etiology, Reward pathway, Reward system, Rhesus, Ribonucleic acid, RMSSD, RNA, RNAi, RNA interference, ROI, Root Mean Square of the Successive Differences, RR internal, SDRR, sE selectin, sE-selectin, Seraton, Seraton*, Serotonin, sICAM 1, sICAM-1, SNS, Soluble intercellular adhesion molecule-1, Standard deviation of RR intervals, Stress blood, Stress Physiological, Stress Physiological/blood, "Stress, Physiological/blood*," Striatum, substantia nigra, subthalamic nucleus, Sympathetic nervous system, Sympathetic NS, Synaptic, Telomere, Triglycerides, Ventral pallidum, VmPFC, Weight

Supplemental Table 4c. Full key terms for search method: Conditions

Key Terms	Academic problem, Acculturation problem, Acute stress disorder, ADD, ADHD, Adjustment, Adjustment disorder, "Adjustment disorder, unspecified, ""Adjustment disorder, with anxiety, ""Adjustment disorder, with depressed mood, ""Adjustment disorder, with disturbance of conduct ,""Adjustment disorder, with mixed anxiety and depressed mood, ""Adjustment disorder, with mixed disturbance of emotions and conduct ,"Adjustment disorders, Adult antisocial behavior, Adverse effects of medication NOS, Age-related cognitive decline, Agoraphobia, Agoraphobia without history of panic disorder, Alcohol abuse, Alcohol dependence, Alcohol intoxication, Alcohol intoxication delirium, Alcohol use disorder, Alcohol withdrawal, Alcohol withdrawal delirium, Amnestic disorder due to another medical condition, Amnestic disorder NOS, Amphetamine, Amphetamine Abuse, Amphetamine Dependence, Amphetamine Intoxication, Amphetamine Intoxication Delirium, Amphetamine Withdrawal , Anorexia, Anorexia nervosa, Antisocial personality disorder, Anxiety, Anxiety disorder, Anxiety disorder due to another medical condition, Anxiety disorder NOS, Asperger, Asperger's disorder, Aspergers, Attention deficit disorder, Attention deficit hyperactivity disorder, Attention-deficit/hyperactivity disorder, "Attention-deficit/hyperactivity disorder, combined type ,""Attention-deficit/hyperactivity disorder, NOS ,""Attention-deficit/hyperactivity disorder, predominantly hyperactive-impulsive type ,""Attention-deficit/hyperactivity disorder, predominantly inattentive type ,"Autism, Autism spectrum, Autism spectrum disorder, Autistic disorder, Avoidant personality disorder, Avoidant/restrictive food intake disorder, BD, Bereavement, Binge eating disorder, Binge-eating disorder, Bipolar, Bipolar and related disorder due to another medical condition, Bipolar disorder, Bipolar disorder NOS, Bipolar I disorder, "Bipolar I disorder, most recent episode depressed ,""Bipolar I disorder, most recent episode hypomanic ,""Bipolar I disorder, most recent episode manic ,""Bipolar I disorder, most recent episode mixed ,""Bipolar I disorder, most recent episode unspecified ,""Bipolar I disorder, single manic episode ,"Bipolar II disorder, Body dysmorphic disorder, Borderline intellectual functioning, Borderline personality disorder, Breathing-related sleep disorder, Brief psychotic disorder, Bulimia, Bulimia nervosa, Caffeine intoxication, Caffeine withdrawal, Cannabis abuse, Cannabis dependence, Cannabis intoxication, Cannabis intoxication delirium, Cannabis use disorder, Cannabis withdrawal, Catatonia associated with another medical disorder, Catatonia disorder due to another medical condition, Central sleep apnea, Child or adolescent antisocial behavior, Childhood disintegrative disorder, Childhood onset fluency disorder , Chronic motor or vocal tic disorder, Circadian rhythm sleep disorder, "Circadian rhythm sleep disorder, delayed sleep phase type ,""Circadian rhythm sleep disorder, jet lag type ,""Circadian rhythm sleep disorder, shift work type,""Circadian rhythm sleep disorder, unspecified type ,"Circadian rhythm sleep-wake disorders, Cocaine abuse, Cocaine dependence, Cocaine intoxication, Cocaine intoxication delirium, Cognitive disorder NOS, Communication disorder, Communication disorder NOS, Conduct disorder, "Conduct disorder, adolescent-

onset type ,""Conduct disorder, childhood-onset type ,""Conduct disorder, unspecified type ,"Conversion disorder, Cyclothymi, Cyclothymi*, Cyclothymia, Cyclothymic, Cyclothymic disorder , Delayed Ejaculation, Delirium, Delirium due to medical condition, Delirium NOS, Delusional disorder, Dementia due to another medical condition, Dementia due to Creutzfeld-Jakob disease, Dementia due to head trauma, Dementia due to HIV disease, Dementia due to Huntington's disease, Dementia due to Parkinson's disease, Dementia due to Pick's disease, Dementia NOS, "Dementia of the Alzheimer's Type, with early onset ,""Dementia of the Alzheimer's Type, with late onset ,"Dependent personality disorder, Depersonalization, Depersonalization disorder, Depersonalization/derealization disorder, Depressed, Depression, Depressive, Depressive disorder, Depressive disorder due to another medical condition, Depressive disorder NOS, Derealization, Derealization disorder, Developmental coordination disorder, Developmental delay, Disinhibited social engagement disorder, "Disorder of infancy, childhood, or adolescence NOS, "Disorder of written expression, Disruptive behavior disorder NOS, Disruptive mood dysregulation disorder, Dissociative amnesia, Dissociative disorder, Dissociative disorder NOS, Dissociative fugue, Dissociative identity disorder, Dyspareunia (not due to a general medical condition), Dyssomnia NOS, Dysthymi, Dysthymi*, Dysthymia , Dysthymic disorder, Eating disorder NOS, Enopresis, "Enopresis, with constipation and overflow incontinence ,""Enopresis, without constipation and overflow incontinence ,"Enuresis, Enuresis (not due to a general medical condition), Erectile disorder, Excoriation disorder, Exhibitionism, Exhibitionistic disorder, Expressive language disorder, Externalizing, Externalizing disorder, Factitious disorder, Factitious disorder NOS, Factitious disorder with combined psychological and physical signs and symptoms, Factitious disorder with predominantly physical signs and symptoms, Factitious disorder with predominantly psychological signs and symptoms, Feeding disorder of infancy or early childhood, Female dyspareunia due to a general medical condition, Female hypoactive sexual desire disorder due to a general medical condition, Female orgasmic disorder, Female sexual interest/arousal disorder, Fetishism, Fetishistic disorder, Frotteurism, Frotteuristic disorder, GAD, Gambling disorder, Gender dysphoria, Gender identity disorder in adolescents or adults, Gender identity disorder in children, Gender identity disorder NOS, General personality disorder, Generalized anxiety disorder, Genito-palvic pain/penetration disorder, Global developmental delay, Hallucinogen abuse, Hallucinogen dependence, Hallucinogen intoxication, Hallucinogen intoxication delirium, Hallucinogen persistng perception disorder, Histrionic personality disorder, Hoarding disorder, Hypercondriasis, Hypersomnia related to axis I or axis II disorder, Hypersomnolence, Hypersomnolence disorder, Hypoactive sexual desire disorder, Hypochondriasis, Identity problem, Illness anxiety disorder, Impulse-control disorder NOS, Inhalent abuse, Inhalent intoxication, Inhalent intoxication delirium, Inhalent use disorder, Inhärent dependence, Insomnia, Insomnia disorder, Insomnia related to axis I or axis II disorder, Intellectual disability, Intermittent explosive disorder, Internalizing, Internalizing disorder, Kleptomania, Language disorder, Learning disorder NOS, Major depressive disorder, "Major depressive disorder, recurrent ,""Major depressive disorder, single episode ,"Major neurocognitive disorder, Major or mild frontotemporal neurocognitive disorder, Major or mild neurocognitive disorder due to alzheimer's disease, Major or mild neurocognitive disorder due to another medical condition, Major or mild neurocognitive disorder due to HIV infection, Major or mild neurocognitive disorder due to huntington's disease, Major or mild neurocognitive disorder due to multiple etiologies, Major or mild neurocognitive disorder due to parkinson's disease, Major or mild neurocognitive disorder due to prion disease, Major or mild neurocognitive disorder due to traumatic brain injury, Major or mild neurocognitive disorder with lewy bodies, Major or mild vascular neurocognitive disorder, Male dyspareunia due to a general medical condition, Male erectile disorder, Male erectile disorder due to a general medical condition, Male hypoactive sexual desire disorder, Male orgasmic disorder, Malingering, Manic, Manic depression, Manic depressive , Manic-depressive , Mathematics disorder, MDD, Medication-Induced Movement Disorder NOS, Medication-Induced Postural Tremor, Mental disorder , Mental disorder NOS due to another medical condition, Mental disorders, Mental illness, "Mental retardation, severity unspecified ,"Mild mental retardation, Mild neurocognitive disorder, Mixed receptive-expressive language disorder, Moderate mental retardation, Mood disorder, Mood disorder due to a general medical condition, Mood disorder NOS, Mood disorders, Narcissistic personality disorder, Narcolepsy, Neglect of child, Neuroleptic Malignant Syndrome , Neuroleptic-Induced Acute Akathisia, Neuroleptic-Induced Acute Dystonia, Neuroleptic-Induced Parkinsonism, Neuroleptic-Induced Tardive Dyskinesia, Nicotine dependence, Nicotine withdrawal, Nightmare disorder, Non-rapid eye movement sleep arousal disorder, Obsessive compulsive disorder, Obsessive compulsive personality disorder, Obsessive-compulsive and related disorder due to another medical condition, Obsessive-compulsive disorder, Obsessive-compulsive personality disorder, Obstructive sleep apnea hypopnea, Occupational Problem, OCD, Opioid abuse, Opioid dependence, Opioid intoxication, Opioid intoxication delirium, Opioid use disorder, Opioid withdrawal, Oppositional defiant disorder, Other (or unknown) substance intoxication, Other (or unknown) substance use disorder, Other (or unknown) substance withdrawal, Other (or unknown) substance-induced disorders, Other alcohol-induced disorders, Other caffeine-induced disorders, Other cannabis-induced disorders, Other female sexual dysfunction due to another medical condition, Other hallucinogen intoxication, Other hallucinogen intoxication, Other hallucinogen use disorder, Other hallucinogen use disorder, Other hallucinogen-induced disorder, Other inhalant-induced disorders, Other male sexual dysfunction due to another medical condition, Other opioid-induced disorders, Other phencyclidine-induced disorders, "Other sedative-, hypnotic-, or anxiolytic-induced disorders ,"Other specified anxiety disorder, Other specified attention-deficit/hyperactivity disorder, Other specified bipolar and related disorder, Other specified delirium, Other specified depressive disorder, "Other specified disruptive, impulse-control, and conduct disorder ,"Other specified dissociative disorder, Other specified elimination disorder, Other specified feeding or eating disorder, Other specified gender dysphoria, Other specified hypersomnolence disorder, Other specified insomnia disorder, Other specified mental disorder, Other specified mental disorder due to another medical condition, Other specified neurodevelopment disorder, Other specified obsessive-compulsive and related disorder, Other specified paraphilia disorder, Other specified personality disorder, Other specified schizophrenia spectrum and other psychotic disorder, Other specified sexual disfunction, Other specified sleep-wake disorder, Other specified somatic symptoms and related disorder, Other specified tic disorder, Other specified trauma- and stress-related disorder, Other stimulant-induced disorders, Other substance abuse, Other substance dependence, Other substance intoxication, Other substance withdrawal, Other tobacco-induced disorders, Pain disorder associated with both psychological factors and a general medical condition, Pain disorder associated with psychological factors, Panic, Panic attack, Panic disorder, Panic disorder with agoraphobia, Panic disorder without agoraphobia, Paranoid personality disorder, Paraphilia NOS, Parasomnia , Parasomnia NOS, Parent-child relational problem, Partner-relational problem, Pathological gambling, Pedophilia, Pedophilic disorder, Persistent (chronic) motor or vocal tic disorder, Persistent depressive disorder, Personality change due to another medical condition, Personality change NOS, Pervasive developmental disorder NOS, Phase of life problem, Phencyclidine abuse, Phencyclidine dependence, Phencyclidine intoxication, Phencyclidine intoxication delirium, Phencyclidine use disorder, Phobia, Phonological disorder, Physical abuse of adult (if by partner), Physical abuse of adult (if by person other than partner), Physical abuse of adult (if focus of attention is on victim), Physical abuse of child, Physical abuse of child (if focus of attention is on victim), Pica, PMDD, Polysubstance dependence, Post traumatic stress disorder, Post-traumatic stress disorder, Posttraumatic stress disorder, Premature (early) ejaculation, Premature ejaculation, Premenstrual dysphoric disorder, Premensturral dysphoric disorder, Primary hypersomnia, Primary insomnia, Profound mental retardation, Provisional tic disorder, Psychological factors affecting medical condition, Psychopathology, Psychotic disorder due to another medical condition, Psychotic disorder due to another medical condition with delusions, Psychotic disorder due to another medical condition with hallucinations, Psychotic disorder NOS, PTSD, Pyromania, Rapid eye movement sleep behavior disorder, Reactive attachment disorder, Reactive attachment disorder of infancy or early childhood, Reading disorder, Relational problem NOS, Relational problem related to a mental disorder or general medical condition, Religious or spiritual problem, Restless legs syndrome, Rett's disorder, Rumination disorder, SAD, Schizoaffective disorder, Schizoid personality disorder, Schizophreni, Schizophreni*, Schizophrenia, Schizophrenia catatonic type, Schizophrenia disorganized type, Schizophrenia paranoid type, Schizophrenia residual type, Schizophrenia undifferentiated type, Schizophreniform disorder, Schizotypal (personality) disorder, Schizotypal personality disorder, Seasonal affective , Seasonal affective disorder, Seasonal affective disorder, Seasonal depression, "Sedadative, hypnotic, or anxiolitic abuse ,""Sedadative, hypnotic, or anxiolitic dependance ,""Sedadative, hypnotic, or anxiolitic intoxication ,""Sedadative, hypnotic, or anxiolitic intoxication delirium ,""Sedadative, hypnotic, or anxiolitic use disorder ,""Sedadative, hypnotic, or anxiolitic withdrawal ,""Sedadative, hypnotic, or anxiolitic withdrawal delirium ,"Selective mutism, Separation anxiety disorder, Severe mental retardation, Sexual abuse of adult (if by partner), Sexual abuse of adult (if by person other than partner), Sexual abuse of adult (if focus of attention is on victim), Sexual abuse of child, Sexual abuse of child (if focus of attention is on victim), Sexual aversion disorder, Sexual disorder NOS, Sexual dysfunction NOS, Sexual masochism, Sexual masochism disorder, Sexual sadism, Sexual sadism disorder, Shared psychotic disorder, Sibling related problem, Sleep disorder due to another general medical condition, Sleep disorder due to hypersomnia type, Sleep disorder due to insomnia type, Sleep disorder due to mixed type, Sleep disorder due to parasomnia type, Sleep eating, Sleep

	talking, Sleep terror, Sleep terror disorder, Sleep walking, Sleep walking disorder, Sleep-related hypoventilation, Social (pragmatic) communication disorder, Social anxiety , Social anxiety disorder, Social phobia, Somatic symptom disorder, Somatization disorder, Somatization disorder NOS, Specific learning disorder, Specific phobia, Speech sound disorder, Stereotypic movement disorder, Stimulant abuse, Stimulant intoxication, Stimulant use disorder, Stimulant withdrawal, Stuttering, Substance abuse, Substance abuse disorder, Substance use, Substance use disorder, Substance use disorders, Substance-abuse, Substance-abuse disorder, Substance-induced disorder, Substance-use, Substance-use disorder, Substance-use disorders, Substance/medication-induced anxiety disorder, Substance/medication-induced bipolar and related disorder, Substance/medication-induced depressive disorder, Substance/medication-induced major or mild neurocognitive disorder, Substance/medication-induced obsessive-compulsive disorder, Substance/medication-induced psychotic disorder, Substance/medication-induced sexual dysfunction, Substance/medication-induced sleep disorder, Tic , Tic disorder, Tic disorder NOS, Tobacco abuse, Tobacco use disorder, Tobacco withdrawal, Tourettes's disorder, Transient tic disorder, Transvestic disorder, Transvestic fetishism, Trichotillomania, Undifferentiated somataform disorder, Unipolar depression, Unspecified alcohol-related disorder, Unspecified anxiety disorder, Unspecified attention-deficit/hyperactivity disorder, Unspecified bipolar and related disorder, Unspecified caffeine-related disorder, Unspecified cannabis-related disorder, Unspecified catatonia, Unspecified communication disorder, Unspecified delirium, Unspecified depressive disorder, "Unspecified disruptive, impulse-control, and conduct disorder ,"Unspecified dissociative disorder, Unspecified elimination disorder, Unspecified feeding or eating disorder, Unspecified gender dysmorphia, Unspecified hallucinogen-related disorder, Unspecified hypersomnolence disorder, Unspecified inhalant-related disorder, Unspecified insomnia disorder, Unspecified intellectual disability, Unspecified mental disorder, Unspecified mental disorder (non psychotic), Unspecified mental disorder due to another medical condition, Unspecified neurocognitive disorder, Unspecified neurodevelopmental disorder, Unspecified obsessive-compulsive and related disorder, Unspecified opioid-related disorder, Unspecified other (or unknown) substance-related disorder, Unspecified paraphilia disorder, Unspecified phencyclidine-related disorder, Unspecified schizophrenia spectrum and other psychotic disorder, "Unspecified sedative-, hypnotic, or anxiolytic-related disorder ,"Unspecified sexual disfunction, Unspecified sleep-wake disorder, Unspecified somatic symptom and related disorder, Unspecified stimulant-related disorder, Unspecified tic disorder, Unspecified tobacco-related disorder, Unspecified trauma-and stress-related disorder, Vaginismus (not due to a general medical condition), Vascular dementia, Vascular dementia uncomplicated, Vascular dementia with delirium, Vascular dementia with delusions, Vascular dementia with depressed mood, Voyeurism, Voyeuristic disorder

Supplemental Table 4d. Full key terms for search method: Stressors

Key Terms	Abandon, Abandoned, Abandonment, Abuse, Abused, Abus*, Accident, Accident or crash, Accidental, Accidental burning, ACE, Adolescence, Adolescent, Adult Survivors of Child Adverse Events, Adverse, Advers*, Adverse childhood event, Adverse childhood events, Adverse childhood experience, Adverse childhood experiences, Adversity, Animal attack, Assault, Assault*, At risk, At-risk, Babies, Baby, Battered child syndrome, Beaten, Being over scheduled, Bereavement, Blight, Body image, Bullied, Bully, Burning, Breakup, Caregiver reported inadequacy of family income, Caregiver-reported inadequacy of family income, Child, Childhood, Children, Child Abandoned, Child Abandon, Child Abuse, Child Adverse Events, Child Foster , Child Orphan, Child Orphaned, Child Syndrome, Child* abuse, Chronic* stress*, Crash, Cyberbully, Cyber-bully, Death, Depriv, Depriv*, Descendant, Difficulty, Difficulty with school work, Disability, Disaster, Discipline, Discrimination, Disrupted home, Distress, Divorce, Domestic violence, Drowning, Drown*, Early adversity, Early life, Early life stress, Early years, Early-life, Earthquake, Economic hardship, ED, ED Visit, ELS, Emergency department, Emergency department visit, Emergency room, Emotional abuse, Emotional needs unmet, Emotional neglect, Environment, Exposure to violence, Family conflict, Family discord, Father substance use, Father substance use disorder, Feeling pressured to behave beyond their ability, Feeling pressured to perform beyond their ability, Feeling pressured to perform or behave beyond their ability, Flood, Grief, Hardship*, Harsh, Harsh parenting, Homeless Youth, Hospitalization, "Hospitalization, ED visit, or invasive medical procedures ,"Illness, Immigrant, Impoverished, Impoverish, Impoverishment, Incarceration, Increased pressure at home, Increased responsibility at home, Infancy , Infant, Injury, Intergenerational trauma, Invasive medical procedures, Invasive medical procedures, Juvenile, Kid, Kidnapped, Kids with parents with cancer, Killed, Lack of attention, Life change events, Life stress, Loss, Loss of job, Low-income, Maltreat*, Maltreatment, Man made disaster, Man-made disaster, Maternal depression, Maternal deprivation, Maternal separation, Maternal fetal relations, Maternal-fetal relations, Minor, Mistreat*, Molest, Molest*, Molestation, Mother treated violently, Mother substance use disorder, Mother substance use, Mother/father/parent Substance use disorder, Natural disaster, Near drowning, Neglect, Neglect*, Neighborhood blight, Neighborhood disorder, Neighborhood risk, Neonatal, Neonate, Newborn, Offspring, Orphan, Over scheduled, Parent in jail, Parent in prison, Parent loss, Parent substance use, Parent substance use disorder, Parent*, Parental death, Parental loss, Parental mental illness, Parental separation, Parents with kids with cancer, Paternal Deprivation , Patient, Peer rejection, Physical abuse, Physical needs unmet, Physical neglect, Poor, Poverty, Poverty Areas, Pregnancy, "Prenatal, hurricane ,"Primary caregiver education level, Primary caregiver education level < high school, Primary caregiver education level < high school graduation, Primary caregiver single parent, Primary caregiver single parenthood, Primary caregiver unemployed, Primary caregiver unemployment, Problem with alcohol, Problem with alcohol or drugs, Problem with drugs, Psychological distress, Psychological stress, Psychological trauma, Psychotrauma, Puberty, Rape, Rape victim, Raped, Receipt of Assistance, Refugee, Separation, SES, Sexual abuse, Sexual assault, Shot, Sick, Single-parent, Social Environment, Socioeconomic status, Socio-economic status, Starting a new school, Stigma, Stress, Stressor, "Stress, ,""Stress, Physiopathology ,""Stress, Psychological ,""Stress, Psychological/physiopathology* ,"Substance use, Suicide, Survivor, Teen, Teenager, Teen*, Toddler, Tragedy, Traged*, Trauma, Trauma*, Traumatology, Upbringing, Victim of violence, Violence, Violent, Violen*, War, Witness another person being beaten, "Witness another person being beaten, raped, threatened with serious harm, shot at, seriously wounded, or killed ,"Witness another person being killed, Witness another person being raped, Witness another person being seriously wounded, Witness another person being shot, Witness another person being shot at, Witness another person being threatened, Witness another person being threatened with serious harm, Witness another person being wounded, Witness violence, Wounded, Young, Youth

References

1. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
2. Beltagy, I., Lo, K. & Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *ArXiv190310676 Cs* (2019).
3. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs* (2019).
4. Wolf, T. *et al.* HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv191003771 Cs* (2020).