# The psycho-env corpus: research articles annotated for knowledge discovery on correlating mental diseases and environmental factors

**Hui Wang[1], Quan Sun[2], Anika Oellrich[3], Honghan Wu[3] and Richard Dobson[3]**
[1] Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom
[2] Department of Informatics, King's College London, United Kingdom
[3] Department of Biostatistics and Medical Informatics, King's College London, United Kingdom
{hui.1.wang, quan.sun, anika.oellrich, honghan.wu, richard.j.dobson}@kcl.ac.uk

## Abstract

While the published scientific literature is used in a biomedical context such as building gene networks for disease gene discovery, it seems to be an undervalued resource with respect to mental illnesses. It has been rarely explored for the purpose of gaining psychopathology insights. This limits our capability of better understanding the underlying mechanisms of mental disorders. In this paper we describe the psycho-env corpus, which aims at annotating published studies for facilitating knowledge discovery on pathologies of mental diseases. Specifically, this corpus focuses on the correlations between mental diseases and environmental factors. We report the first preliminary work of psycho-env on annotating 20 articles about two mental illnesses (bipolar disorder and depression) and two particular environmental factors - light and sunlight. The corpus is available at `https://github.com/KHP-Informatics/psycho-env`.

## 1 Introduction

The success stories of cognitive computing (e.g., IBM Watson's Jeopardy game) and deep learning (e.g., DeepMind's AlphaGo) have sparked a massive wave of using artificial intelligence (AI) to improve numerous aspects of our daily life. Not surprisingly, healthcare is among the hottest areas. For example, IBM Watson is now utilised in decision support for lung cancer at the Memorial Sloan Kettering Cancer Center. However, AI models require data to derive better understanding of the underlying mechanisms of diseases before they can really improve existing treatments or increase the recovery rate. Unfortunately, the lack of data is a major hurdle in many areas of the clinical domain, such as understanding the pathologies of mental illnesses.

As with other diseases, it has been established that mental illnesses are influenced in their origins and pathology by environmental factors. For example, it has been found that higher rates of schizophrenia occur in people of Caribbean origin than general population living in the UK [Fung *et al.*, 2006]. To date, no complete list of environmental factors for all existing mental illnesses has been compiled that can be used for patient screening and planning treatment strategies [Rutter, 2005].

While the published scientific literature is used in a biomedical context such as building gene networks for disease gene discovery [Lage *et al.*, 2007] or symptom networks of inheritable human disorders [Zhou *et al.*, 2014], it seems to be an undervalued resource with respect to mental illnesses. It has been rarely explored for the purpose of gaining psychopathology insights. The potential of this resource lies within the amount and variety of data available: all journals that publish scientific results are covered mostly since 1966, though some even date back to 1809. Although there is a body of work trying to identify "extended" phenotypes [Oellrich *et al.*, 2016; Groza *et al.*, 2015; Collier *et al.*, 2015], however, none of these efforts included environmental factors, which are necessary to understand gene-phenotype relationships. In order to make use of this tremendous resource for finding potential environmental factors that (i) cause, (ii) contribute to and (iii) influence the origin and pathology of mental illnesses, (AI backed) automated methods are needed to digest the large quantities of existing data.

In order to facilitate this endeavour, data collection and annotation would be required to identify relevant studies and the representation of environmental factors in the published literature. In this paper we describe the psycho-env corpus[1], which is a manually curated dataset from the abstracts of 20 published studies on associations between two mental illnesses (bipolar disorder and depression) and one particular environmental factor - light. We believe this is the first effort to produce curated corpus for knowledge discovery on associations between mental illness and environmental factors.

In the next section, we introduce the article selection, annotation process, annotation tool used and data format of annotations. In section 3, we describe the psycho-env corpus and discuss the limitation of this work. Finally, we conclude our work in section 4.

[1] `https://github.com/KHP-Informatics/psycho-env`

## 2 Materials and methods

### 2.1 Article selection

In this preliminary study, we limited our scope on two types of mental disorders (i.e., bipolar and depression) and one particular environmental factor - light (including sunlight and light in general). A manual retrieval method was adopted to search and select articles from various bibliographic databases and search engines. This was to ensure that we were able to identify the most relevant and representative investigations in this domain for the pilot study. The search and selection process are briefly described in the following.

#### Literature search

The bibliographic databases and search engines used were MEDLINE (accessed via PubMed search engine), Web of Science and Google Scholar. The aim was to look for relevant and representative research articles including clinical studies, case reports and clinical trials published during the period from May 1877 to May 2017.

The terms used for searching disorders included: *bipolar, manic* and *depression*, while terms for environmental factors included *sunlight, "light therapy"* and *phototherapy*. In some situations, extra constrains were added to narrow down the search results, e.g., *clinical trial, case reports* and etc.

In general, we found PubMed combined with Google Scholar can produce the most comprehensive list for our searches. For example, when searching *sunlight and bipolar disorder*, PubMed results contained 7 relevant hits, Google Scholar had 6, and Web of Science gave 5. All combined, there were 8 distinct relevant hits. The overlap between PubMed and Google Scholar was 5 - PubMed brought in 2 new results and Google Scholar added 1, while all results from Web of Science were covered by other two search services.

Also, we found the terminologies used in the literature are quite heterogenous. For example, when denoting the usage of light in the therapy, many different terms were used - *bright-light therapy*, *light therapy* and *phototherapy*. Therefore, we found it necessary to follow the reference graph of articles to check and include more articles or search terms.

#### Article selection

The studies were selected based on the following inclusion criteria:

- published as an original article in a peer-reviewed journal
- designed as a clinical trial, pilot study or case report
- used light or sunlight as one of the investigation aspects or treatment alternatives
- subjects were diagnosed as bipolar disorder or depression

Table 1 contains the distribution of the psycho-env articles.

Table 1: Articles in psycho-env corpus

| Mental disorders and light factor | Articles |
|---|---|
| Sunlight to bipolar disorder | 7 |
| Light to bipolar disorder | 5 |
| Sunlight to depression | 4 |
| Light to depression | 4 |

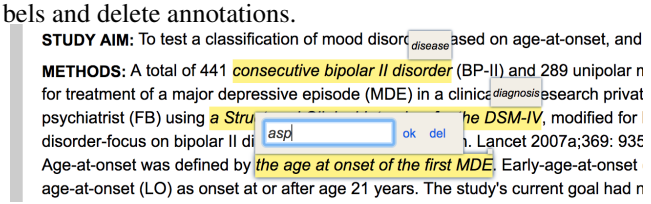
### 2.2 Annotation guidelines and process

When reviewing the articles, curators were asked to extract the following information to create a correlation between mental illnesses and environmental factors. When combined together, the annotated items should be able to a) capture the most important aspects for deriving the correlations and b) form a concise description of the study. For well-defined clinical concepts like disorders, phenotypes and clinical measurements, the curators were asked to map them to UMLS (Unified Medical Language System)[2] concepts using a UMLS search tool.

1. The most important finding(s) of the study (e.g., *Bipolar inpatients in E rooms (exposed to direct sunlight in the morning) had a mean 3.67-day shorter hospital stay than patients in W rooms* [Benedetti *et al.*, 2001]).
2. Environmental factors. Although this preliminary study focused on light only, other types of environmental factors might need to be annotated as well because they were used in the study to derive or measure light factors, such as "latitudes 6.3 to 63.4 degrees from the equator". Type of environmental factors including, but not limited to: sunlight exposure, seasonal pattern, sunlight in springtime, natural light, 36 collection sites from 23 countries, and monthly climate variables.
3. Environmental factor classification or measurement. This type of information includes the conceptual classification or quantity metrics for environmental factors investigated in the study, such as meteorological data on light intensity, the amount of sunlight exposure (i.e. insolation), maximum monthly increase in solar insolation and etc.
4. Mental disorders. As mentioned earlier, two types of diseases were to be curated in this work: bipolar and depression disorders. Any diseases that are specific types of these diseases need to be annotated, which include, but not limited to, bipolar I disorder, recurrent depression, non-seasonal depression, and rapid cycling bipolar.
5. Investigation aspects of disorders - the aspects of disease pathologies or phenotypes that were investigated in the study, such as the onset of bipolar disorder, mood swings, length of hospitalization and plasma melatonin levels.
6. Diagnosis methods (if available), such as Young Mania Rating Scale (YMRS).
7. Patient cohort information including number of patients, patient demographic information, and control/case settings.

[2] https://www.nlm.nih.gov/research/umls/

Figure 1: PsychoEnv Annotator User Interface: yellow highlights are the annotated texts; grey popups are labels (types) of the highlights; The popup dialog allows to add/change labels and delete annotations.




8. Data collection methods and data sources, such as patient records and/or direct interviews and NASA Surface Meteorology and Solar Energy (SSE) database.
9. Data analysis methodologies, such as Autoregressive Integrated Moving Average (ARIMA) method.

To the best of our knowledge, this is the first attempt to curate literature in this particular domain. A large part of the curation is unknown to us, for example, what aspects of diseases were studied and how they were quantified, what terminologies were used to describe both clinical and environmental concepts, how environmental factors were measured and etc. Considering this underdeveloped nature, we adopted an agile curation process, which was designed to be adaptive and able to achieve continuous improvement. The idea was borrowed from the agile software development. Technically, articles were partitioned into several subsets and curations were conducted on each subset at a time. After each curation step, a curator meetup would be arranged to discuss problems encountered and the lessons learned, and subsequently propose amendments on the curation guidelines for improving the next rounds. We found this iterative process and efficient inter-curator communications very helpful and effective.

### 2.3 Annotation tool and annotation data format

A browser based annotation tool, PsychoEnv annotator, was used for annotating articles. The tool is backed with an automated article highlighting service described in [Wu *et al.*, 2017]. PsychoEnv annotator is available on Github: `https://github.com/KHP-Informatics/psycho-env`. Figure 1 is a screenshot of PsychoEnv annotator being used for annotating a PubMed article. Features of the tool include:

- Easy to setup: the annotation tool is a Chrome extension and the backend service is cloud based. Any article with an online XHTML version (e.g., PubMed article abstracts) is available for annotating immediately without the need of any preprocessing.
- Easy to use: all curation operations are browser based, which minimises the learning curve of curation process. In addition, the free text labelling allows project-wise acronyms, which speeds up the process.
- Easy to share: associating annotations with web-addressed articles makes the annotations directly retrievable either for the browser visualisation by using PsychoEnv annotator or for software agents by RESTful API calls.

Table 2: Annotation data format

| | |
|---|---|
| Article URI | The web URL of the article's web version, e.g. `https://www.ncbi.nlm.nih.gov/pubmed/24953482` |
| Annotation node locator | The locator is composed of two components:<br>1. a jQuery[3] selector that locates the parent element of the text node, where the annotation appears;<br>2. an integer number that indicates the index of the text node within its parent's children list.<br>For example: a locator can be { Selector: *ABSTRACTTEXT:eq(1)*, Index: *0* } |
| Annotation offsets | The offsets have two integer components: start_offset and end_offset, where start_offset indicates the start position of the annotated text in its annotation node's text content and end_offset indicates the end position. |
| Text | The text content of the annotation |
| Type | The type of the annotation |

- Structure preserving: compared to most existing annotation tools, PsychoEnv annotator is featured by its unique capability of locating annotations on the XHTML DOM tree of the articles' web pages (see annotation node locator in table 2). This associates the annotations with semi-structured DOM trees and, in turn, brings these tree structures as additional and easy-to-consume features to software models.

The annotation data format is a 5-element tuple as described in Table 2.

## 3 Results and discussion

### 3.1 Corpus description

The psycho-env corpus resulted in 27 annotated text nodes that mark mental disorder mentions, 30 annotated text nodes that mark environmental factors, 25 annotated text nodes that mark environmental factor classifications/measurements and 23 annotated sentences marked as important findings. These numbers are summarized in Table 3 which also shows the average number of annotations and range of annotations per article in the 20 articles in the corpus.

The psycho-env corpus was selected to represent bipolar and depression disorders associated with two environmental factors - sunlight and (general) light. The aim was to have a similar coverage on each of the four sub-domains (as shown in Table 1) so that we could cover relatively diverse topics within a preliminary study. We summarised the major types of annotations in table 4. Duplicated instances have been removed using a syntax approach - string comparison . The first observation is that the environmental concepts seem to be very heterogeneous (1.4 per article for light factors and 2.05 per article for light measurements) even when we limited the scope on light only. However, a close inspection on

Table 3: Three major annotation types; averages were computed over the set of articles that contained that annotation type.

| Type | # anns | # articles | Avg. |
|---|---|---|---|
| Mental disorders | 39 | 20 | 1.95 |
| Environmental factors | 33 | 19 | 1.73 |
| Environmental classification / measurement | 43 | 18 | 2.38 |

Table 4: Major annotation types and their distinct instance numbers.

| Annotation Type | Number |
|---|---|
| Mental disorders | 31 |
| Disorder phenotypes | 21 |
| Phenotype measurements | 17 |
| Diagnosis criteria | 7 |
| Environmental factors (Light) | 28 |
| Environmental factors (Other) | 6 |
| Light/Sunlight classification / measurement | 41 |
| Analysis methodologies | 7 |

the list of instances revealed that many different terms might mean the same concepts. This suggests the necessity of having a consistent terminology so that different mentions of the same instances can be mapped. The second interesting observation is that the numbers of specific disorders, phenotypes and their measurements are relative large considering only 2 disorders were selected. This suggests very little overlaps between studies, which might imply that the curation could be very efficient in terms delivering new knowledge.

### 3.2 Discussion

The main purpose of this preliminary study is to conduct a small scale case study on limited types of mental illnesses and environmental factors. Therefore, the number of documents annotated is rather small. But it has resulted with a very valuable experience, which gave us a good understanding about the quality and representation of environmental factors and their associations with mental disorders. Particularly, the typed annotations as summarised in table 4 can be used to populate controlled vocabularies or ontologies to represent knowledge in this domain.

The corpus covers four subdomains of associations of mental disorders and environmental factors as depicted in table 1. The authors are confident that they have covered the most representative studies in the top 3 subdomains. However, regarding the last subdomain - Light to depression, due to a relatively large body of available studies, the selected four articles might not cover the most representative studies.

## 4 Conclusion

In order to facilitate knowledge discovery on the pathologies of mental disorders, we initiated work on psycho-env corpus, which is dedicated to curating the associations between mental illnesses and environmental factors from published literature. The first version reported in this paper focused on bipolar and depression disorders associated with lights, and was curated from abstracts of 20 articles. Both the annotation tool and the corpus are open source and publicly available.

## Acknowledgments

The work was supported by NIHR Biomedical Research Centre for Mental Health, the Biomedical Research Unit for Dementia at the South London, the Maudsley NHS Foundation Trust and Kings College London, and European Union's Horizon 2020 research and innovation programme under grant agreement No 644753(KConnect).

## References

[Benedetti *et al.*, 2001] Francesco Benedetti, Cristina Colombo, Barbara Barbini, Euridice Campori, and Enrico Smeraldi. Morning sunlight reduces length of hospitalization in bipolar depression. *Journal of affective disorders*, 62(3):221–223, 2001.

[Collier *et al.*, 2015] Nigel Collier, Anika Oellrich, and Tudor Groza. Concept selection for phenotypes and diseases using learn to rank. *Journal of biomedical semantics*, 6(1):24, 2015.

[Fung *et al.*, 2006] WL Alan Fung, Dinesh Bhugra, and Peter B Jones. Ethnicity and mental health: the example of schizophrenia in migrant populations across europe. *Psychiatry*, 5(11):396–401, 2006.

[Groza *et al.*, 2015] Tudor Groza, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M Couto, Gareth Baynam, Andreas Zankl, and Peter N Robinson. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, 2015:bav005, 2015.

[Lage *et al.*, 2007] Kasper Lage, E Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, 25(3):309–316, 2007.

[Oellrich *et al.*, 2016] Anika Oellrich, Nigel Collier, Tudor Groza, Dietrich Rebholz-Schuhmann, Nigam Shah, Olivier Bodenreider, Mary Regina Boland, Ivo Georgiev, Hongfang Liu, Kevin Livingston, et al. The digital revolution in phenotyping. *Briefings in bioinformatics*, 17(5):819–830, 2016.

[Rutter, 2005] Michael Rutter. How the environment affects mental health. *The British Journal of Psychiatry*, 186(1):4–6, 2005.

[Wu *et al.*, 2017] Honghan Wu, Anika Oellrich, Christine Girges, Bernard de Bono, Tim J.P. Hubbard, and Richard J.B. Dobson. Automated PDF highlighting to support faster curation of literature for Parkinson's and Alzheimer's disease. *Database*, 2017(1):bax027, 2017.

[Zhou *et al.*, 2014] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature communications*, 5, 2014.