*** Reliability and validity analyses for the coding of information entered into the Ehealth4MDD database ***

Author: Franziska Burger, Mark A. Neerincx, Willem-Paul Brinkman
Interactive Intelligence Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology

Corresponding Author: Franziska Burger

Contact Information: f.v.burger@tudelft.nl

****

*** General introduction ***

This dataset contains all data and analysis scripts pertaining to the research conducted for the CyberPsychology23 conference paper: “EHealth4MDD: a database of e-health systems for the prevention and treatment of depressive disorders.” In the scope of the research conducted and described in this paper, we have developed a relational database to systematically describe e-mental health systems for the prevention and treatment of Major Depressive Disorder (MDD). For the purpose of creating this database, literature had to be retrieved from PubMed, Scopus, and Web of Knowledge and filtered for inclusion and exclusion based on title, abstract, and full-paper. Samples of records at each stage were double coded. Once the final body of literature was identified, information from the papers had to be extracted (coded) and entered into the database. Four of the database attributes were selected to be double coded again on samples. Furthermore, a set of scales was developed of which we assessed concurrent validity. We here deliver:

1. eHealth4MDD database: the version of the database as it was used in the analyses conducted for the paper
2. reliability and validity analyses: all information pertaining to the assessment of reliability of the main rater and the assessment of the validity of the five scales including:
    - -instructions for second raters
    - -materials provided to the second raters to do the rating task
    - -ratings as assigned by the first rater
    - -ratings as assigned by the second raters
    - -analysis scripts to determine reliability and validity

****

*** Methodological information ***

The data was collected by selecting and coding published research on e-mental health systems for the treatment and prevention of major depressive disorder. This

was conducted by coder 1 or C1. The literature screening process to identify the research as well as four high-inference variables were assessed for reliability of C1. This was done by asking second coders (C2 for the literature screening and one of the high-inference variables, and C3 for the remaining three high-inference variables) to look at samples of these coding variables.
Furthermore, five scales were developed to assign a degree of technological sophistication to the functions of the e-mental health systems. To assess concurrent validity of the scales, five separate coders (C4-8) were provided with a set of items for a specific scale to code on the basis of their intuitive (naive) understanding of "technological sophistication." After some time passed, the same coders were again asked to code the same items, now with knowledge of the scale levels (informed). Naive and informed scores were compared to determine intra-rater agreement as a measure of concurrent validity. The informed scores were also correlated with the scores assigned by C1 to all items for interrater agreement.
Lastly, to assess whether the five scales to measure "technological sophistication" for different types of items are comparable, one coder, again C2, received all items that C4-8 had received but mixed, i.e. not separated by type. He was also asked to first score the items naively and then provided with all the scales to assign informed scores.

****

*** Data specific information ***

Details the folders, files, and column headers of tabular data included in the dataset. The R scripts in the preprocessing_and_raw_data folder produce the *.txt files required for the full_analysis_CYPSY23.R script.

A. Database

1. *ehealth4mdd_cypsy.sql*: mysql database used in data collection and preprocessing. More details concerning the database can found at http://insyprojects.ewi.tudelft.nl:8888/.
2. *database_content.html*: for those interested in all data stored in the database including codes and their descriptions, this html files includes this information and can be opened using any browser.

B. Analyses

1. analysis

contains the script and all preprocessed data for running the reliability and validity analyses of which the results are reported in the two tables of the paper.

1.1 full_analysis_CYPSY23.R

# Analysis script producing the numbers that can be found back in the tables of the paper, printing all these results to the output_CYPSY23_tables.txt file

1.2 output_CYPSY23_tables.txt

# Output file produced by the R script containing all results of the analysis

1.3 function_classification_C1C3.txt

# File with ID of instantiation of a function (brief description taken from articles) and the function key that coder 1 (C1) and coder 3 (C3) assigned to this instantiation
#
# data fields
#
#1 VersionKey: unique identifier for each ehealth system version used in this reliability check
#2 FunctionKey: unique identifier for a specific function that was found to be in version by either C1 or C3
#3 C1: 1 if C1 assigned this function, na if C1 did not assign this function
#4 C3: 1 if C3 assigned this function, na if C3 did not assign this function

1.4 function_identification_C1C3.txt

# File with ID of instantiation of a function (brief description taken from articles) and the function key that C1 and C3 assigned to this instantiation
#
# data fields
#
#1 ID: unique identifier for each function instantiation
#2 C1: the function key of the corresponding function as assigned by C1
#3 C3: the function key of the corresponding function as assigned by C3

1.5 function_therapy_mapping_C1C3.txt

# File with mapping of therapy intervention class (e.g. CBT, CT, ACT, psychodynamic psychotherapy, ...) to functions
#
# data fields
#
#1 FunctionKey: unique identifier for each intervention function of an ehealth system
#2 Description: a brief description of the functionality provided by each function
#3 C1: the respective code for a therapy intervention class assigned to the function by C1
#4 C3: the list of codes of therapy intervention classes assigned to the function by C3

1.6 literature_screening_C1C2.txt

# Contains the results of the title, abstract, and full paper filtering of records as done by coder 1 (C1) and coder 2 (C2)
#
# data fields
#
# PaperKey: key of the record as assigned by endnote on computer of C1
# C1: coding assigned by coder 1
# C2: coding assigned by coder 2

     # Stage: which stage of the filtering process the record was evaluated at (either title, abstract, or full paper stage)

1.7 validation_C1CX.txt

     # File with the assigned ehdts scores of coders 2, and 4-8 when they did not know what each number of the scale meant, i.e. when they were asked to

     # rate function instantiations according to their intuition of technological sophistication, when coders did know the definitions of scale levels,
     # and the assigned scores by C1 to the same items")
     #
     # data fields
     #
     #1 VersionKey: unique identifier for the Version
     #2 FunctionKey: unique identifier for the function, together with the VersionKey this identifies the item that was rated
     #3 CXnoscale: eHDTS score as assigned by CX (where X is element of {4,5,6,7,8}) without knowing the definitions of the scale levels
     #4 CXscale: eHDTS score as assigned by CX when knowing the definitions of the scale levels
     #5 C1scale: eHDTS score as assigned by C1 when knowing the definitions of the scale levels
     #6 coderID: ID of the coder that completed the coding (e.g. 4 -> C4)

2. data_collection

Contains task descriptions and instructions for the second raters as well as the materials that raters needed to complete the task.

2.1 function_classification_and_identification
contains the files that were used by C3 and task descriptions for C3 to identify functions in papers and classify function instantiations as specific functions
2.1.1 *coding_file_classification_empty.xlsx*
     # an empty excel document for C3 to write down the results of the classification process
     #
     # data fields
     #
     #1 ID: unique identifier for the function instantiation
     #2 Description: textual description of the function instantiation as taken from the original publication
     #3 Function?: an empty column for C3 to enter the assigned code
2.1.2 *coding_file_identification_empty.xlsx*
     # an empty excel document for C3 to write down the results of the identification process
     #
     # data fields
     #
     #1 ArticleID: unique identifier for the article
     #2 Title: article title
     #2 Functions?: an empty column for C3 to enter all identified functions

2.1.3 *Coding Instructions.docx*: instructions for C3 for the identification and classification tasks
2.1.3 *function_list.txt*

> # a reference list for C3 with all the possible functions
> #
> # data fields
> #
> #1 FunctionKey: unique identifier of the function
> #2 Function: textual description of the function
> #3 FunctionType: whether the function is an intervention or a support function, and if it's a support functions, whether it's planning, execution, monitoring, or social support
> #4 Therapy: which therapeutic framework the function is linked to, codes are explained in the file.

2.2 function_therapy
contains the files that were used by C3 and task descriptions for C3 to map functions to their therapeutic frameworks
2.2.1 *coding_file_therapy_mapping_empty.xlsx*

> # an empty excel document for C3 to write down the results of mapping intervention functions to therapeutic intervention classes.
> #
> # data fields
> #
> #1 FunctionKey: unique identifier of the function
> #2 Function: textual description of the function
> #3 Therapy: an empty field for C3 to note down which therapeutic framework(s) she thought the function would best fit with.
> #4 Comments: an empty field for C3 to note down thoughts while assigning therapeutic framework(s)

2.2.2 Coding_Instructions.pdf: instructions and task descriptions provided to C3 for linking functions to therapies.
2.2.2 *list_of_therapies.txt*: a reference list of the possible therapeutic intervention classes for C3 to choose from, linking the respective code to the name of the therapy
2.2.3 *therapy_dataset_creation.R*: R script writing the information from the functions table in the database to a text file

2.3 literature_screening
contains the publication IDs that were selected to train C2 in the literature screening process as well as the final publication IDs that were selected for double coding by C2
2.3.1 *filter_results_coder1.txt*: the codings of C1 for each stage of the literature screening process and all records.

> # Contains the results of the title, abstract, and full paper filtering of records as done by coder 1
> #
> # data fields
> #
> #1 PaperKey: key of the record as assigned by endnote
> #2 Title.R1: filtering information of first author on title (1-include, 0-exclude)

#3 Abstract.R1: filtering information of first author on abstract (1-include, 0-exclude, NA-was excluded on title already)
#4 Full.R1: filtering information of first author on full paper (1-include, 0-exclude, NA-was excluded on title or abstract already)

2.3.2 *filtering_sampling.R*: R script to sample records for title, abstract, and full-paper stage of the screening process.

2.3.3 *f*inal_samples: contains the record numbers (as assigned by EndNote to the records) for all three stages of the sampling process for both training the second coder and for testing.

2.3.3.1 test: final samples for testing
2.3.3.1.1 *abstract_codes.txt*
2.3.3.1.2 *full_codes.txt*
2.3.3.1.3 *title_codes.txt*
2.3.3.2 train: final samples for training
2.3.3.2.1 *abstract_codes.txt*
2.3.3.2.2 *full_codes.txt*
2.3.3.2.3 *title_codes.txt*

2.4 validation_cards
contains files for the five different sets of cards that were used in the scale validation study. These were provided to raters in the form of a card sorting task.

2.4.1 *execution_cards.txt*
#the execution support items used in the validation study
#
# data fields
#
#1 label: labels the item with [*VersionKey*, *FunctionKey*][Component: *FunctionName*][Platform: *ICT Technology*]
#2 description: textual description of the function instantiation as it appears in the specific version.

2.4.2 *intervention_cards.txt*: the intervention function instantiations used in the validation study
#the intervention function instantiations used in the validation study
#
# data fields
#
#1 label: labels the item with [*VersionKey*, *FunctionKey*][Component: *FunctionName*][Platform: *ICT Technology*]
#2 description: textual description of the function instantiation as it appears in the specific version.

2.4.3 *monitoring_cards.txt*: the monitoring support function instantiations used in the validation study
#the monitoring support function instantiations used in the validation study
#
# data fields
#
#1 label: labels the item with [*VersionKey*, *FunctionKey*][Component: *FunctionName*][Platform: *ICT Technology*]
#2 description: textual description of the function instantiation as it appears in the specific version.

2.4.4 *planning_cards.txt*: the planning support function instantiations used in the validation study

```
#the planning support function instantiations used in the validation study
#
# data fields
#
#1 label: labels the item with [*VersionKey*, *FunctionKey*][Component: *FunctionName*][Platform: *ICT Technology*]
#2 description: textual description of the function instantiation as it appears in the specific version.
```

2.4.5 *social_cards.txt*: the social support function instantiations used in the validation study

```
#the social support function instantiations used in the validation study
#
# data fields
#
#1 label: labels the item with [*VersionKey*, *FunctionKey*][Component: *FunctionName*][Platform: *ICT Technology*]
#2 description: textual description of the function instantiation as it appears in the specific version.
```

3. preprocessing_and_raw_data

3.1 function_classification

contains the raw data files needed to produce the data file for the reliability analysis of the classification of function instantiations as being a certain function

3.1.1 *classifying_C1.txt*

```
# File with ID of instantiation of a function (brief description taken from articles) and the function key that C1 assigned to this instantiation
#
# data fields
#
#1 ID: unique identifier for each function instantiation
#2 FunctionKey: the function key of the corresponding function
```

3.1.2 *classifying_C3.txt*

```
# File with ID of instantiation of a function (brief description taken from articles) and the function key that C3 assigned to this instantiation
#
# data fields
#
#1 ID: unique identifier for each function instantiation
#2 FunctionKey: the function key of the corresponding function
```

3.1.3 *classifying_preprocessing.R*: R script that merges the above two txt files taking into consideration the key mapping

3.1.4 *key_mapping.txt*

```
# since there was an ordering according to function type in the IDs of C1, this ordering was randomized to obtain new IDs for C3. Key_mapping.txt contains
# the mapping of old keys (C1) to new keys (C3)
#
```

# data fields
#
#1 Order_Final: the new ordering of instantiations
#2 Order_Initial: the original ordering of instantiations

3.2 function_identifying
3.2.1 *identifying_C3.txt*
# Contains the splitting up of systems/versions into functions as done by C3
#
# Data Fields:
#
#1 Articles: unique identifier for an article in which a version of an ehealth system is described
#2 Functions: a list of the functions that C3 identified to be in the system described in the article
3.2.2 *identifying_preprocessing.R*: R script that merges the identified functions per system of C3 with those of C1 (taken from the database)

3.3 function_therapy_mapping
3.3.1 *function_therapy_C1C3.txt*
# Contains the results of the mapping of functions to therapeutic frameworks as conducted by C3 merged with the codes assigned by C1
#
# data fields
#
#1 FunctionKey: unique identifier of the function
#2 Description: textual description of the function
#3 Therapy.C1: the therapeutic intervention framework that had been assigned to this function by C1
#4 Therapy.C3: the therapeutic intervention framework that had been assigned to this function by C3
#5 Comments: comments made in the file by C3 while coding
#6 Discussion: points of discussion when iterating over the unclear ones together.
3.3.2 *function_therapy_C3_raw.xlsx*: the file with the mappings of intervention functions to their therapy classes as returned by C3, it has the same columns as the function_therapy_C1C3.txt file only without the merged column of C1 codes.
3.3.3 *mapping_preprocessing.R*: R script transforming the codings of C3 into ones that are compatible with C1 (process of coding was different for the two coders)

3.4 literature_screening
3.4.1 *abstract_exclude_dpl.txt*
# contains a list of all the records that were excluded by the first coder based on abstract, exported from endnote, contains some duplicate records
#
# data fields
#
#1 PaperKey: key of the document assigned by endnote upon import
#2 Filter: filtering information of first author (ee - exclude on title, ie - exclude on abstract, ii - include on abstract)

#3 Author: authors of record
#4 Year: year when record was published
#5 Title: title of record

3.4.2 *abstract_include_dpl.txt*

# contains a list of all the records that were included by the first coder based on abstract, exported from endnote, contains some duplicate records
#
# data fields
#
#1 PaperKey: key of the document assigned by endnote upon import
#2 Filter: filtering information of first author (ee - exclude on title, ie - exclude on abstract, ii - include on abstract)
#3 Author: authors of record
#4 Year: year when record was published
#5 Title: title of record

3.4.3 *abstract_test_final_C2.txt*

# Contains codings of records that coder 2 (C2) coded on abstract
#
# data fields
#
#1 Column1: record key
#2 Column2: coding of C2

3.4.4 *all_refs_no_dpl.txt*: contains all records without duplicates, exported from endnote

# Contains all records without duplicates, exported from endnote
#
# data fields
#
#1 PaperKey: key of the document assigned by endnote upon import
#2 Author: authors of record
#3 Year: year when record was published
#4 Title: title of record

3.4.5 *filtering_preprocessing.R*: script that merges the different files to create one file containing all codes of C1 and C2 for the different filtering stages of literature screening

3.4.6 *full_test_final_C2.txt*: contains codings of records that C2 coded on full paper

# Contains codings of records that coder 2 (C2) coded on the full paper
#
# data fields
#
#1 Column1: record key
#2 Column2: coding of C2

3.4.7 *ii_papers_key_mapping.txt*

# File for manual editing of papers included at title and abstract stage
#
# data fields
#
#1 Title: title of the paper
#2 PaperKeyDB: PaperKey as assigned in the eHealth4MDD database
#3 Exclude: decision of whether to include or exclude the paper as taken by C1

#4 PaperKey: PaperKey of the record as assigned by EndNote

3.4.8 *ii_papers_key_mapping_edited.txt*: manually edited file (some codes could not be matched properly)

# manually edited file (some codes could not be matched properly on title)
#
# data fields
#
#1 Title: title of the paper
#2 PaperKeyDB: PaperKey as assigned in the eHealth4MDD database
#3 Exclude: decision of whether to include or exclude the paper as taken by C1
#4 PaperKey: PaperKey of the record as assigned by EndNote

3.4.9 *papers_from_db.txt*

# Contains all records that were included based on abstract by first coder as exported from the ehealth4mdd database
#
# data fields
#
#1 PaperKeyDB: key of the document assigned by database
#2 Year: year when record was published
#3 Title: title of record
#4 Database: origin of record (1- PubMed, 2-Scopus, 3-WebOfScience, 4-manually added), 4 has no corresponding key in endnote
#5 Exclude: 1 if excluded by first coder on the basis of the full paper, NA if included

3.4.10 *title_test_final_C2.txt*: contains codings of records that C2 coded on title

# Contains codings of records that coder 2 (C2) coded on title
#
# data fields
#
#1 Column1: record key
#2 Column2: coding of C2

3.4.11 *title_exclude_dpl.txt*

# contains a list of all the records that were excluded by the first coder based on title, exported from endnote, contains some duplicate records#
# data fields
#
#1 PaperKey: key of the document assigned by endnote upon import
#2 Filter: filtering information of first author (ee - exclude on title, ie - exclude on abstract, ii - include on abstract)
#3 Author: authors of record
#4 Year: year when record was published
#5 Title: title of record

3.5 validation

3.5.1 raw_data: contains exported files from the card sorting

3.5.2 *validation_preprocessing.R*: R script to combine the results from the card sorting into one data frame with eHDTS ratings of C1 and CX for all function types

****
****

****

*** Software requirements for running the analyses ***

Data selection, preprocessing, and analyses were conducted with R version 3.5.0. The necessary libraries are detailed in each script.

The version of the database that was used is included as an .sql file. The database is a MySQL database and can be run on a local MySQL server (in my case, version 14.14 Distrib 5.7.22 for osx10.13) for the sake of replicating the preprocessing and data selection steps of the studies.

Set the working directory to the location of the CYPSY23_analyses folder initially (in my case: setwd("~/surfdrive/CYPSY23_eHealth4MDD_analyses/Analyses/”)). All R scripts assume this working directory.

****
****
****