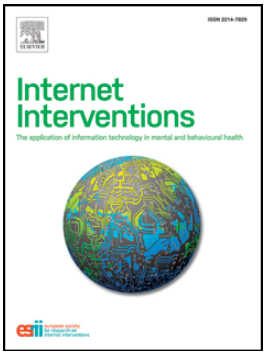


# Using multi-relational data mining to discriminate blended therapy efficiency on patients based on log data

Artur Rocha, Rui Camacho, Jeroen Ruwaard, Heleen Riper

PII: S2214-7829(17)30076-3
DOI: doi:10.1016/j.invent.2018.03.003
Reference: INVENT 196

To appear in: *Internet Interventions*

Received date: 4 August 2017
Revised date: 28 February 2018
Accepted date: 6 March 2018

Please cite this article as: Rocha, Artur, Camacho, Rui, Ruwaard, Jeroen, Riper, Heleen, Using multi-relational data mining to discriminate blended therapy efficiency on patients based on log data, *Internet Interventions* (2018), doi:10.1016/j.invent.2018.03.003

# Using multi-relational data mining to discriminate blended therapy efficiency on patients based on log data

Artur Rocha[a], Rui Camacho[b], Jeroen Ruwaard[c], Heleen Riper[c]

[a] *Centre for Information Systems and Computer Graphics, INESC TEC Porto, Portugal*
[b] *DEI & Faculdade de Engenharia & LIAAD-INESC TEC, Universidade do Porto, Portugal*
[c] *Vrije Universiteit Amsterdam, Department of Clinical Psychology De Boelelaan 1081, 1081 HV Amsterdam*

**Abstract**

**Introduction:** Clinical trials of blended Internet-based treatments deliver a wealth of data from various sources, such as self-report questionnaires, diagnostic interviews, treatment platform log files and Ecological Momentary Assessments (EMA). Mining these complex data for clinically relevant patterns is a daunting task for which no definitive best method exists. In this paper, we explore the expressive power of the multi-relational Inductive Logic Programming (ILP) data mining approach, using combined trial data of the EU E-COMPARED depression trial.

**Methods:** We explored the capability of ILP to handle and combine (implicit) multiple relationships in the E-COMPARED data. This data set has the following features that favor ILP analysis: 1) Time reasoning is involved; 2) there is a reasonable amount of explicit useful relations to be analyzed; 3) ILP is capable of building comprehensible models that might be perceived as putative explanations by domain experts; 4) both numerical and statistical models may coexist within ILP models if necessary. In our analyses, we focused on scores of the PHQ-8 self-report questionnaire (which taps depressive symptom severity), and on EMA of mood and various other clinically relevant factors. Both measures were administered during treatment, which lasted between 9 to 16 weeks.

*URL:* artur.rocha@inesctec.pt (Artur Rocha), rcamacho@fe.up.pt (Rui Camacho)

**Results:** E-COMPARED trial data revealed different individual improvement patterns: PHQ-8 scores suggested that some individuals improved quickly during the first weeks of the treatment, while others improved at a (much) slower pace, or not at all. Combining self-reported Ecological Momentary Assessments (EMA), PHQ-8 scores and log data about the usage of the ICT4D platform in the context of blended care, we set out to unveil possible causes for these different trajectories.
**Discussion:** This work complements other studies into alternative data mining approaches to E-COMPARED trial data analysis, which are all aimed to identify clinically meaningful predictors of system use and treatment outcome. Strengths and limitations of the ILP approach given this objective will be discussed.

*Keywords:* Multi-relational data mining, Internet intervention, Moodbuster, Log data, Ecological momentary assessment

## 1. Introduction

E-COMPARED [7] is an European multi-centre trial, comparing face-to-face cognitive behavior therapy (cbt) for depression with so-called blended cbt, in which treatment is provided through a mix of face-to-face and online contact[1]. At 5 out of 8 research sites of E-COMPARED (DE, FR, NL, UK & PL), the on-line component of the blended treatment was delivered through an internet-based system called Moodbuster/ICT4D [19][13], through which detailed logs of system usage were systematically collected, as well as Ecological Momentary Assessments (EMA) [14].

The wealth of data collected in trials such as E-COMPARED may be key in the development of personalized treatments. To understand the different ways in which patients use the system and how this translates to better (or worse) outcomes, detailed patient-level data such as collected in the E-COMPARED trial are crucial. At present, however, it is not clear what analytic techniques are most suitable to identify relevant patterns in the data. Mining these complex data for clinically relevant patterns is a daunting task for which, at present, no 'silver bullet' method exists.

This paper describes the use of Inductive Logic Programming (ILP; [9]), a data mining method that focuses strongly on implicit relationships between

[1] `http://www.e-compared.eu`

multi-relational data. In our models, we set out to identify links between factors such as the number of hours spent in on-line therapy, the number of messages exchanged with the therapist, or the number of EMA ratings, to explain observed differences in treatment outcomes.

## 2. Inductive Logic Programming Framework

Inductive Logic Programming (ILP) [12] is a flavor of multi-relational learning [4] and, therefore, also a sub-area of Machine Learning (ML). ILP is in the intersection of Statistics and Logic Programming (LP). From LP it inherits the the representation scheme for both data and models — a subset of First Order Logic (FOL). It is a supervised ML method. ILP addresses the problem of inducing hypotheses (as predicate definitions) from examples and background knowledge. According to [8], an ILP learner requires an initial theory BK (background knowledge) and evidence E (examples), to induce a theory H (hypothesis) that, together with BK, *explains* some properties of E.

In traditional ILP, E comes in two forms: positive and negative . In this setting ILP is applied to binary classification problems. Positive examples (E+) are instances of the concept to learn, whereas negative examples are not. Negative examples are used to avoid over generalization.

Due to the use of FOL to encode both data and BK, structured data can be easily handled in ILP. The third ingredient, H, is also encoded in FOL and can therefore represent highly complex models. Traditional ILP systems transform the induction process into a search over a very large space (sometimes infinite) which may cause efficiency problems when dealing with complex problems. To address this problem, the user may constrain the language of H and use a set of parameters for that purpose.

ILP's logic representation of the induced models has been used to leverage several scientific applications. Although, in general, ILP does not perform much better that "propositional" learners like Support Vector Machines, Decision Trees, etc, the goal of its use is to get comprehensible models that contribute to explain the phenomena that produced the data.

ILP systems like Aleph [15], April [5] or Indlog [2] have a highly powerful expressive language to verify the constraints mentioned above to constrain the hypothesis' language. In our study we have used the Aleph [15] ILP system that implements the MDIE (see [10]) method to induce hypotheses.

## 3. Related work

Other data analysis and prediction works have been performed in the scope of E-COMPARED (see e.g. [18], [1] and [17]), mostly focused in the short term prediction of important traits for depression such as mood. Some of these works have made use of data like the one used for the BK in this experiment to attempt to improve the accuracy of the aforementioned models, but they did not focus on trying to understand which factors had impact on the outcome of the therapy.

ILP has relevant applications to problems in complex domains like natural language and molecular computational biology [11]. An initial and major work were ILP was used in a scientific setting with the goal of building understandable models was in a rational drug design domain were ILP was used to predict the mutagenicity of a set of drugs [16]. The data set had been previously studied using a statistics approach [3], with good predictive results but no clue was provided to understand why the drugs were mutagenic. The ILP algorithm identified a set of substructures that the domain experts found relevant for the explanation of mutagenicity.

## 4. Experiments

### *4.1. Data*

For this experiment we considered a dataset containing anonymous log data from 201 patients from different test sites using the ICT4D/Moodbuster platform in 5 different countries.

Planned treatment durations were different across research sites. In DE, PL, and the UK, treatment was delivered to patients who were recruited in primary care, in a scheduled treatment duration of approximately 6 to 13 weeks. In FR and NL, patients were recruited in specialized treatment settings. At these sites, the scheduled treatment duration was 16 to 20 weeks.

In spite of these differences (treatments lasted between 9 to 16 weeks), sites followed a common trial protocol, which makes results comparable. Even though most patients improved along the treatment, it was observed that some patients performed particularly well, achieving significant improvement before the expected treatment duration. Other patients did not recover as much, even when considering the full extent of the therapy.

Considering that improvement takes into account the difference between the final and initial scores, records with less than two PHQ-8 assessments

were purged, resulting in a final data set comprising data for 179 unique patients. Table 1 summarizes base data for this experiment considering participants with at least 2 PHQ-8 evaluations.

| COUNTRY | Participants | Weblog events | Messages | EMA Ratings | PHQ assessments |
|---|---|---|---|---|---|
| DE | 82 | 23613 | 1861 | 20779 | 439 |
| FR | 23 | 8907 | 214 | 3915 | 102 |
| NL | 25 | 5479 | 250 | 3989 | 88 |
| PL | 26 | 7309 | 236 | 4256 | 120 |
| UK | 23 | 4103 | 30 | 2359 | 103 |
| Total N | 179 | 49411 | 2591 | 35298 | 852 |

Table 1: Number of: participants per country, Weblog events, Messages, PHQ assessments

The features in Table 2, derived from the log data and EMA datasets, have been used to write ILP background knowledge — set of predicates to encode the ILP model.

*4.2. Data Preparation*

In order to setup the ILP model, a set of positive and negative examples (E) needs to be provided. For this setting we will use the PHQ-8 score as an index accepted by experts to measure the severity of depression. Therefore, if we consider the difference between the final and the initial PHQ-8 scores a measure for the outcome of the therapy, we need to define a cutoff value to distinguish successful from unsuccessful interventions.

A simplistic approach is to consider a positive outcome a reduction of at least 1 point in the PHQ-8 score. However, most clinicians will disregard this approach as an effective measure of outcome since it does not take into account the error inherent to the associated evaluation instrument (the PHQ-8 questionnaire). As a rule of thumb, several practitioners will consider an intervention successful if there is a decrease of at least 50% relative to the first PHQ score.

It is possible, however, to determine a clinical significant change using statistical methods[6]. According to this method a reliable change index (RC) of at least 1.96 means that the post-test score ($x_f$) is likely to reflect a

| | Predicates derived from EMA and Log Data | Description |
|---|---|---|
| Metadata & EMA | • Total EMA ratings | Total EMA ratings by type, per patient |
| | • EMA mood ratings adherence | Average of the patient's daily mood ratings (Total EMA mood ratings/Last day in therapy when ratings were received) |
| | • Nationality | Country of origin of the patient |
| | • Number of treatment days | Total number of days in the treatment |
| Exchanged MSG | • Messages sent by therapist | Number of messages that have been sent by the therapist |
| | • Length of messages sent by therapist | Number of characters contained in the messages sent by the therapist |
| | • Messages sent by patient | Number of messages that have been sent by the patient |
| | • Length of messages sent by patient | The number of characters contained in the messages sent by the patient |
| Moodbuster Usage LOG | • Number of Web sessions | Number of distinct on-line visits to the Moodbuster Web site during treatment |
| | • Total time spent on-line | Sum of the time spent in all the page visits to the Moodbuster Web site during treatment |
| | • Number of distinct pages visited | Number of distinct therapy pages viewed in the course of the on-line treatment |
| | • Number of on-line exercises complete | Number of on-line exercises complete in the course of treatment, including repetitions |
| | • Number of on-line modules complete | Number of therapy modules complete during the on-line treatment |

Table 2: Background knowledge.

real change from a pretest score ($x_i$), where:

$$RC = \frac{x_f - x_i}{S_{diff}}$$

Therefore we can calculate the standard error $SE = S_{dev}\sqrt{1 - r_{xx}}$, where $S_{dev}$ is the standard deviation for the initial PHQ-8 score in the whole population ($S_{dev} = 4.48$) and $r_{xx}$ is the test-retest reliably score for the ques-

tionnaire[2] ($r_{xx} = 0.85$), calculating the value of $S_{diff}$ defined as $\sqrt{2(SE)^2}$. Then, in order to obtain an RC index of at least 1.96, the minimum change ($x_f - x_i$) with clinical significance is 4.81. In this case, a negative change is sought.

Last but not least, we can define a cutoff for patients that experienced a significant change and finished their treatments with mild to no depression — final PHQ8 inferior to 10.

In summary, the following cutoffs have been defined for this experiment:

- **Positive change:** patients that improved at least one point in the PHQ-8 score ($x_f - x_i \geq -1$);
- **50% Improvement:** patients whose final PHQ-8 score is at least 50% less than the initial one ($x_f \leq \frac{x_i}{2}$);
- **Significant change:** patients that improved at least 5 points in the PHQ-8 scale ($x_f - x_i > -4.81$);
- **Clinically Improved:** patients that improved at least 5 points and finished with less than 10 in the PHQ-8 scale ($x_f - x_i > -4.81; x_f < 10$).

According to the defined cutoffs, Table 3 summarizes patient improvements by country and in total.

| Country | Participants (at least 2 PHQ-8 scores) | Positive change ($x_f - x_i \geq -1$) | 50% Improvement ($x_f \leq \frac{x_i}{2}$) | Significant change ($x_f - x_i \geq -5$) | Clinically Improved ($x_f - x_i \geq -5; x_f < 10$) |
|---|---|---|---|---|---|
| DE | 82 | 73 (89%) | 38 (46%) | 57 (70%) | 46 (56%) |
| FR | 23 | 19 (83%) | 7 (30%) | 9 (39%) | 7 (30%) |
| NL | 25 | 22 (88%) | 8 (32%) | 13 (52%) | 9 (36%) |
| PL | 26 | 21 (81%) | 14 (54%) | 18 (69%) | 15 (58%) |
| UK | 23 | 19 (83%) | 9 (39%) | 13 (57%) | 9 (39%) |
| Total N | 179 | 154 (86%) | 76 (42%) | 110 (61%) | 86 (48%) |

Table 3: Summary of improvements according to predefined cutoff conditions

[2] http://patienteducation.stanford.edu/research/phq.pdf

*4.3. ILP Experiments*

When using ILP as a data analysis tool we had two goals in mind: i) to induce a comprehensible model that could be interpreted as a plausible explanation for the fast recovering of some of the patients encoded in the data set and; ii) to determine which features of the treatment may have a relevant impact in the recovery speed.

We have divided the Background Knowledge (BK) concerning patients data into three sets: predicates analyzing patient and EMA metadata (further referred to as EMA); predicates analyzing Moodbuster usage logs by patient (LOG); and predicates analyzing the exchanged messages between patients and caretakers (MSG). We then setup seven data analysis studies, using each of the background knowledge subsets separately and combinations between these.

Given time constraints and the heavy ILP run times, instead of running the model for all the cutoff conditions, we have chosen one of the most strict cutoff condition for improvement — *clinically improved*, which seems also the most sensible criteria for clinicians.

In all of the studies the following Aleph parameter's values were set. The hypothesis space was bound to a limit of 2 million clauses. The maximum clause length was set to 14, meaning that a clause could not have more than 14 literal (including the head literal). The sample size was set to 3, meaning that 3 positive examples were used to seed 3 searches across the corresponding hypothesis space and only after the 3 search procedures the best clause found was accepted. With the above parameters set we have carried out a search for a good theory varying two other parameters that have high impact on the theory quality: *minpos*; and *noise*. *minpos* avoids Aleph to accept clauses that "explain" less than *minpos* positive examples. *noise* specifies that an acceptable clause may "explain" *noise* negative examples. We have experimented with the following set of *minpos* percentage values: 5; 6 ; 7; 8; 10; and 15. We have used the following *noise* values: 3; 4; and 5. Models were assessed using a 10-fold Cross Validation (CV) procedure. *Accuracy* and *F-measure* were the metrics used to estimate the model's quality. The total number of Aleph runs was 7 (BK subsets) $\times$ 5 (*minpos* values) $\times$ 3 (*noise* values) $\times$ 10 (CV folds) giving a total of 1050 runs.

*4.4. Results*

Table 4 summarizes the performance of the ILP model according to the most clinically relevant cutoff conditions, used to distinguish between pa-

tients that improved and that did not improve.

| BK subset | Accuracy (%) | Precision (%) | Recall (%) | F-measure (%) |
|---|---|---|---|---|
| LOG | 60 (18) | 58.3 (13.1) | 63.8 (16.8) | 59.9 (12.0) |
| MSG | 65.7(16.6) | 68.1 (21.3) | 57.5 (23.4) | 60.8 (20.0) |
| EMA | 60.2 (15.9) | 60.2(16.9) | 53.5(19.4) | 56.1(17.6) |
| LOG + MSG | 59.7 (7.6) | 60.33(9.8) | 51.4(14.3) | 54.7(10.2) |
| EMA + MSG | 60 .4(13.8) | 60.4(15.6) | 52.5(15.3) | 55.4(14.5) |
| EMA + LOG | 62.1(8.2) | 62.4(11.5) | 52.0(16.7) | 56.0(15.9) |
| EMA + LOG + MSG | 62.5 (10.3) | 59.2(6.2) | 58.0(13.0) | 57.5(6.0) |

Table 4: Performance results for the combination of background subsets of predicates. The pos/neg criteria used was *clinically improved*. The values in the cells are the average and standard deviation (within the parenthesis) of a 10-fold Cross Validation procedure.

The number of patients that improved according to this cutoff condition was 86 out of a total of 179, meaning that 93 patients did not clinically improve (although they may have improved to some extent). This means that the majority class accounts for 52% (93/179) of the patients, which sets a base line for the accuracy of each set of BK.

The precision (a sort of accuracy considering only the positive values) is calculated as the number of true positives (positives correctly predicted by the model) divided by the number of cases the model predicted as positive. The recall is a measure of sensitivity, calculated by dividing the true positives by the total number of effective positives. F-measure is a harmonic mean between precision and recall.

### *4.5. Discussion*

Considering the LOG and EMA BK separately, the accuracy of the model is merely 8% above the majority class. While this a positive result, it is not expressive enough to affirm that the usage of the Internet-based platform and

associated App, collectively designated as Moodbuster, predicts the clinical improvement of the patient *per se*. The predicates for messages exchanged (MSG BK) account for an average of nearly 66% across all cross-validations which seems to suggest that a higher engagement between patient and therapist has a positive effect on the outcome of the therapy. However, combining MSG either with LOG (Web platform adherence) or with EMA (Mobile App adherence) does not improve the performance of the model (60% in both cases).

In addition there is a high standard deviation in the accuracy of all the BK subsets, associated to relatively low values of precision, recall and F-measure. We have to say also that the size of the data set is small if we consider the number of countries of origin of the patients. The representatives of each country in the "test block" of a 10-fold cross validation might be rather small or non existent (Ex: if one has only 5 recovered patients from a country and we have 10 folds , some folds would not have a single representative of recovered patient from that country).

As an example of the understandability of the rules resulting from ILP, we would like to point out two examples:

- Polish patients that spent at least 58.1 minutes online up to the first PHQ assessment and were in treatment for a total of at least 48 days, clinically improved. [Rule covering 12 positive cases out of 15 possible and only 2 negative cases].

- German patients that spent less than 85 minutes online up to the first PHQ assessment and were in treatment for a total of at least 83 days, clinically improved. [Rule covering 12 positive cases out of 46 possible and only 1 negative case].

Seemingly contradictory rules such as the previous ones, associated to the fact that no outstanding rules emerged from the runs (combinations of predicates that explain a high percentage of the results), led us to perform a few summaries on the data to obtain an explanation of why that happened.

The first insight comes from the last column of Table 3, which shows that the number of positive examples according to the chosen cutoff condition is not very high, when divided by country.

Additionally, when looking at Table 5, containing the summary of usage data for the patients that clinically improved (labeled Y) and the ones that

did not (labeled N) — considering the whole dataset (last row of the table), there is a very dim variation in terms of usage between the two groups. There are some large differences when considering individual countries such as FR, NL or UK, which are attenuated by an inverse tendency in the most representative group of the population from DE.

| CN | CI? | Page Visits | Minutes Online | Sessions | Exercises | Distinct Pages | Days from First to Last Visit |
|---|---|---|---|---|---|---|---|
| DE | Y | 263 | 346 | 16 | 114 | 121 | 98 |
| | N | 320 | 405 | 17 | 136 | 118 | 98 |
| FR | Y | 586 | 527 | 27 | 262 | 124 | 163 |
| | N | 300 | 333 | 15 | 122 | 98 | 100 |
| NL | Y | 325 | 317 | 20 | 142 | 104 | 128 |
| | N | 160 | 187 | 9 | 62 | 76 | 86 |
| PL | Y | 292 | 332 | 14 | 109 | 120 | 94 |
| | N | 267 | 411 | 18 | 109 | 94 | 46 |
| UK | Y | 257 | 342 | 13 | 97 | 125 | 117 |
| | N | 128 | 173 | 6 | 45 | 86 | 76 |
| Tot. | Y | 300 | 355 | 17 | 127 | 119 | 108 |
| | N | 254 | 321 | 14 | 104 | 99 | 87 |

Table 5: Summary of Moodbuster platform usage for Clinically Improved (CI) *vs* Not Clinically Improved

In addition to the previous results, Table 6 focuses only on the clinically improved group and provides a summary of patient data when the cutoff condition *clinically improved* was met, in comparison with the full length of the treatment.

In average, when patients could be considered clinically improved, they had already recovered 90% of their total improvement, done around 60% of the modules (3 in 5 modules, excluding introduction) and spent 60% of the total intervention time (less than 8 of 12 weeks).

It is worth mentioning that no big differences were found in terms of therapy performance across the countries. For example, the patients that crossed the CI threshold in DE did so even faster than the average. The high engagement found in the LOG BK for the group that did not clinically

improve, maps to the patients that kept using the platform even though they did not meet the threshold.

| Country | Improvement | | Modules Complete | | Days in Treatment | |
|---|---|---|---|---|---|---|
| | Total | CI | Total | CI | Total | CI |
| DE | 9 | 8 (90%) | 6 | 3 (63%) | 81 | 46 (56%) |
| FR | 10 | 8 (84%) | 5 | 3 (57%) | 119 | 60 (54%) |
| NL | 9 | 7 (91%) | 5 | 3 (69%) | 114 | 68 (63%) |
| PL | 11 | 10 (91%) | 5 | 3 (62%) | 65 | 43 (66%) |
| UK | 9 | 8 (92%) | 6 | 4 (65%) | 84 | 62 (70%) |
| Total N | 9 | 8 (90%) | 5 | 3 (63%) | 85 | 51 (60%) |

Table 6: Patient data when Clinically Improved (CI) *vs* Total Treatment

## 5. Conclusions and Future Work

Insight from the ILP model suggests adherence to the use of the Moodbuster platform (Web, App and messaging) has a positive influence on the outcome of the treatment, even though the numbers are modest: 60% accuracy. The strict conditions of the *clinically improved* cutoff filter out lighter improvements (see Table 3), which can still mean that patients have greatly benefited from the intervention, even if they did not finish the treatment with mild to no depression. This is also consistent with the fact that the data used for this experiment is not final data, so several patients were still in treatment.

Results from the previous section also show that patients classified under the *clinically improved* group not only achieved the treatment goals, but have converged to their goal rather fast — roughly 90% of the total improvement occurred in 60% of the full intervention time. The reasons behind this fast improvement could be further studied with a larger dataset enabling the application of induction only on the clinically improved group.

Other future work includes running the models with the complete datasets for the whole E-COMPARED trials data and using different cutoff criteria, with the aim of unveiling higher accuracy background knowledge (BK). Another possible research direction is fine-tuning BK sets to include only the most promising predicates.

## References

[1] Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., Ruwaard, J., 2016. How to predict mood? delving into features of smartphone-based data.

[2] Camacho, R., 1998. Inducing models of human control skills. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 107–118.
URL http://dx.doi.org/10.1007/BFb0026679

[3] Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J., Hansch, C., 1991. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. Journal of medicinal chemistry 34 (2), 786–797.

[4] Džeroski, S., Lavrač, N. (Eds.), 2001. Relational Data Mining, 1st Edition. Springer-Verlag.

[5] Fonseca, N. A., Silva, F., Camacho, R., 2006. April – An Inductive Logic Programming System. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 481–484.
URL http://dx.doi.org/10.1007/11853886_42

[6] Jacobson, N. S., Truax, P., 1991. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of consulting and clinical psychology 59 (1), 12.

[7] Kleiboer, A., Smit, J., Bosmans, J., Ruwaard, J., Andersson, G., Topooco, N., Berger, T., Krieger, T., Botella, C., Baños, R., et al., 2016. European comparative effectiveness research on blended depression treatment versus treatment-as-usual (e-compared): study protocol for a randomized controlled, non-inferiority trial in eight european countries. Trials 17 (1), 387.

[8] Lavrač, N., Flach, P. A., 2001. An extended transformation approach to inductive logic programming. ACM Transactions on Computational Logic (TOCL) 2 (4), 458–494.

[9] Muggleton, S., 1991. Inductive logic programming. New generation computing 8 (4), 295–318.

[10] Muggleton, S., 1995. Inverse entailment and Progol. New Generation Computing, Special issue on Inductive Logic Programming 13 (3-4), 245–286.

[11] Muggleton, S., 1999. Inductive logic programming: Issues, results and the challenge of learning language in logic. Artificial Intelligence 114 (1), 283 – 296.
URL http://www.sciencedirect.com/science/article/pii/S0004370299000673

[12] Muggleton, S., De Raedt, L., 1994. Inductive logic programming: Theory and methods. JLP 19/20, 629–679.

[13] Rocha, A., Henriques, M. R., Lopes, J. C., Camacho, R., Klein, M., Modena, G., Van de Ven, P., McGovern, E., Tousset, E., Gauthier, T., et al., 2012. Ict4depression: Service oriented architecture applied to the treatment of depression. In: Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on. IEEE, pp. 1–6.

[14] Shiffman, S., Stone, A. A., Hufford, M. R., 2008. Ecological momentary assessment. Annu. Rev. Clin. Psychol. 4, 1–32.

[15] Srinivasan, A., 2001. The aleph manual.

[16] Srinivasan, A., Muggleton, S. H., King, R., Sternberg, M., 1994. Mutagenesis: Ilp experiments in a non-determinate biological domain. In: Proceedings of the 4th International Workshop on Inductive Logic Programming, volume 237 of GMD-Studien. pp. 217–232.

[17] van Breda, W., Hoogendoorn, M., Eiben, A., Andersson, G., Riper, H., Ruwaard, J., Vernmark, K., 2016. A feature representation learning method for temporal datasets. In: Computational Intelligence (SSCI), 2016 IEEE Symposium Series on. IEEE, pp. 1–8.

[18] van Breda, W., Pastor, J., Hoogendoorn, M., Ruwaard, J., Asselbergs, J., Riper, H., 2016. Exploring and comparing machine learning approaches for predicting mood over time. In: Innovation in Medicine and Healthcare 2016. Springer, pp. 37–47.

[19] Warmerdam, L., Riper, H., Klein, M. C., van de Ven, P., Rocha, A., Henriques, M. R., Tousset, E., Silva, H., Andersson, G., Cuijpers, P.,

2012. Innovative ict solutions to improve treatment outcomes for depression: the ict4depression project. Annual Review of Cybertherapy and Telemedicine 181 (1), 339–343.

**Highlights**

- Preliminary data from blended therapy in 5 sites of the E-COMPARED RCT is used
- Multi-relational Inductive Logic Programming (ILP) models were constructed
- Usage data for the Internet-based component of therapy is used for predicates
- Clinically relevant criteria are used to discriminate positive/negative examples
- Models were assessed using a 10-fold cross validation procedure