Check for updates

Applying Data Mining to Develop a Short Version of The Public Health Informatics Competencies for Primary Health Care (PHIC4PHC) Questionnaire

Enny Rachmani [1], Evina Widianawati [1], MG. Catur Yuantari [1], Respati Wulandari [1] and Hanif Pandu S [1,2]

[1] Faculty of Health Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[2] Semarang District Health Office, Semarang Indonesia

enny.rachmani@dsn.dinus.ac.id

Abstract. The successful integration of digital technology in the healthcare sector frequently encounters obstacles due to various factors, mainly when patients are involved, and the other is the competency of health workers. The previous research developed a questionnaire to measure public health informatics competencies (PHIC4PHC) consisting of 4 main categories or domains: cognitive proficiency, technical proficiency, ethical proficiency, and health information literacy. It consisted of 42 questions. There is a need to design a shorter version of the Public Health Informatics Competencies for Public Health Practitioners (PHIC4PHC) assessment tool that can effectively measure capabilities in the field of public health informatics with the same precision. This study extensively describes the methodology employed in developing a condensed version of the PHIC4PHC questionnaire, utilizing feature selection techniques from a data mining approach. The dataset was explored using RapidMiner studio version 9.1. This study utilized an optimization function by employing a feature selection operator in RapidMiner. The study employed a range of characteristics to achieve the highest level of accuracy in the classification. Specifically, 8 to 15 questions were selected as the target experiments for the model. This study proposed a questionnaire called PHIC4PHC-S12Q, which consists of 12 features. The model PHIC4PHC-S12Q was selected due to its fulfillment of the primary indicator and the inclusion of all indicators in ethical proficiency, distinguishing it from other models from experiment results. Future studies need to conduct the confirmatory factor analysis to compare the model construct by data mining technique

Keywords: Primary Health Care, Digital Health Literacy, Questionnaire.

© The Author(s) 2024
A. Khoironi et al. (eds.), *Proceedings of the International Seminar and Workshop Public Health (ISWHOPHA 2023)*, Advances in Health Sciences Research 75,
https://doi.org/10.2991/978-94-6463-421-1_2

1 Introduction

The continual development of digital technology is gradually influencing human behavior. The expansion of digital penetration has already advanced to Digital Industry 4.0 and continues progressing towards Society 5.0. [1, 2] The Covid-19 pandemic has rapidly accelerated digital transformation throughout several aspects of human life, particularly within the healthcare industry.

Primary health care (PHC) has been widely implemented as the primary approach in numerous low- and middle-income countries to attain Universal Health Coverage (UHC) with a focus on equity, patient-centeredness, and comprehensive care. [3] The enhancement of primary healthcare (PHC) is of utmost importance in providing to individuals residing in remote areas, where there is a pressing need for improvements in infrastructure, proficient healthcare personnel, suitable health technologies, financial backing, and comprehensive management of health programs. [4]

Multiple studies have documented the growing utilization of health information technologies in diverse primary healthcare (PHC) initiatives. These technologies encompass inpatient electronic registries, processing and evaluation programs, management systems, clinical decision support systems, surveillance tools, and patient monitoring systems. [5, 6]

The successful integration of digital technology in the healthcare sector frequently encounters obstacles due to various factors, particularly when patients are involved, and the other is the competency of health workers. [7, 8] It is imperative to recognize these shortcomings, particularly within primary health care centers (PHCs). [9] The successful deployment of health information technology (IT) in primary healthcare (PHC) is influenced by various aspects, with different contexts presenting unique obstacles. Among these elements, human resources have been identified as particularly crucial in ensuring success. [6] The competencies of health workers in the field of digital health within Primary Health Care can be classified as Public Health Informatics Competencies. The gaining of these competencies is essential for the successful implementation of digital health technologies in primary healthcare.

The previous research developed a questionnaire to measure public health informatics competencies (PHIC4PHC) consisting of 4 main categories or domains: cognitive proficiency, technical proficiency, ethical proficiency, and health information literacy [10]. This questionnaire has 42 questions and could be categorized into four level competencies. [11] Previous studies have identified challenges in the interviewing process, primarily attributed to time constraints. As a result, there is a need to design a shorter version of the Public Health Informatics Competencies for Public Health Practitioners (PHIC4PHC) assessment tool that can effectively measure capabilities in the field of public health informatics with the same precision. This study extensively describes the methodology employed in developing a condensed version of the PHIC4PHC questionnaire, utilizing feature selection techniques from a data mining approach. Furthermore, an examination of the classification accuracy of the resulting model is conducted. The features were carefully chosen, and the attribute sets were subsequently reduced in size. The impact of these modifications on the classification accuracy was then examined.

2 Methods

2.1 Data Set

This study used data set from previous research of Public Health Informatics Compenecies for Primary Health Care (PHIC4PHC), the PHIC4PHC is the first-ever questionnaire that evaluates the essential abilities needed by primary healthcare (PHC) professionals in the digital health era. These competencies comprise computer skills, ethical skills, and health literacy skills.[10] The data set consist of 581 data with 42 questions.

The PHIC4PHC questionnaire comprised 42 items, each rated on a scale from strongly disagree to strongly agree, with a score range of 1 to 4.

$$index = (mean - 1) * 50/3 \text{ [11]}$$

The scores of all 42 questions were converted to a scale ranging from 0 to 50. In this scale, 0 represents the lowest score and 50 represents the highest score for public health informatics competencies (PHIC). Consequently, thresholds and ranges were established to categorize PHIC4PHC into four levels: basics, literate, fluent, and master.[11] The distribution of competencies was shown in Table 1.

Table 1. Data Set Distribution Frequency of PHIC

Variable	N (%)	Basics N (%)	Literate N (%)	Fluency N (%)	Mastery N (%)
Gender					
Male	98 (16,8)	3 (3,1)	32 (32,7)	45 (45,9)	18 (18,4)
Female	483 (83,1)	25 (5,2)	156 (32,3)	221 (45,8)	81 (16,8)
Education					
Below High School	1 (0,2)	0 (0)	0 (0)	0 (0)	1 (100)
High School	45 (7,7)	4 (8,9)	15 (33,3)	21 (46,7)	5 (11,1)
Vocational	384 (66,1)	19 (4,9)	146 (38)	167 (34,5)	52 (13,5)
Bachelor	146 (25,1)	5 (3,4)	27 (18,5)	75 (51,4)	39 (26,7)
Master/Doctor	5 (0,9)	0 (0)	3 (60)	2 (40)	0 (0)

2.2 Data Pre-processing

The prior study's dataset contained no missing values and verified an uneven composition for dataset labels such as basics (5%), literate (32%), fluent (46%), and master (17%). This study used repeated data to balance the dataset until it reached a balanced composition. The process of balancing was carried out by duplicating data with the basic label nine times, while fluent and literate labels were no-duplicated and mastered twice. The final dataset contained 904 data points that were used to perform the data mining process with a feature selection tool.

2.3 Experiment Setting

The dataset was explored using RapidMiner studio version 9.1 in this study. This study utilized an optimization function by employing a feature selection operator in RapidMiner. Feature selection is a fundamental task in data mining that involves identifying the most important features for classification.

This study examined the precision of classifying PHIC4PHC based on the outcome of feature selection in the experiment. RapidMiner offers a diverse array of search techniques, including evolutionary algorithms. Performance measurement is required for all search methods to assess the potential effectiveness of a feature subset on a given data set. The datasets were validated using a validation function that employed a cross-validation operator to assess the accuracy of the model. The classification model in this study was predicted using a k-NN (k Nearest Neighbor) algorithm. [12]

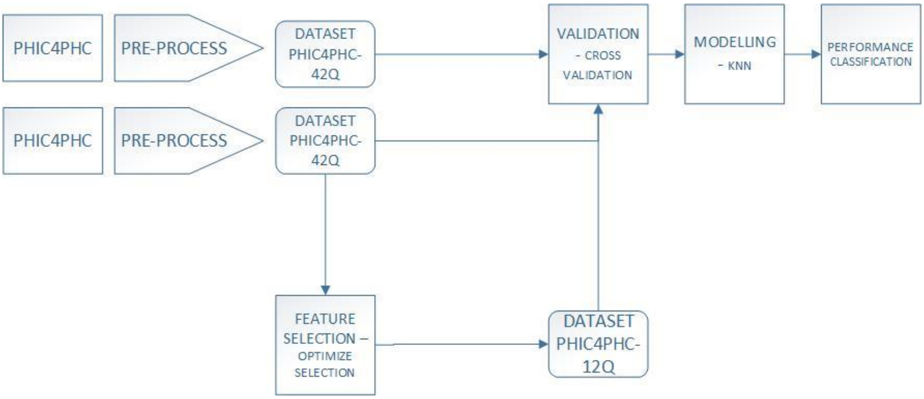

Fig. 1. Experiment process

2.4 Parameter

The current PHIC4PHC comprises a total of 42 qualities. Thus, this study chose 15 attributes as the maximum goal of the concept in a concise PHIC4PHC questionnaire. The study employed a range of characteristics to achieve the highest level of accuracy in the classification. Specifically, 8 to 15 questions were selected as the target experiments for the model. The work employed a genetic algorithm (GA) to maximize the selection of features from the dataset, utilizing a novel selection operator. GA was selected due to its ability to optimize with both continuous and discrete variables, its capability to handle a high number of variables without requiring derivative information, its ability to generate a list of optimal variables, and its compatibility with various data or analytical functions. [13]. The study used k-NN combined with GA to optimize the performance classification [14].

3 Theory

3.1 Genetic Feature Selection Algorithm

The Genetic Algorithm (GA) is a prominent Heuristic Algorithm that emulates the theory of evolution. It is an Evolutionary Algorithm that employs the principles of natural selection and crossover to identify the most effective solution[15]. The search space parameters of a genetic algorithm (GA) are encoded as strings, which are referred to as chromosomes. A group of these strings is referred to as a population. At the beginning, a diverse population is generated, representing various points throughout the search space. Each string is paired with an aim and fitness function that quantifies the quality of the string. Following the principle of natural selection, a small number of strings are chosen and each is allocated a specific number of duplicates that are placed into the breeding pool. Biologically inspired operators, such as crossover and mutation, are used on these strings to produce a new generation of strings. The process of selection, crossover, and mutation persists for a predetermined number of generations or until a termination condition is met. The following code provides the fundamental algorithmic procedures for Genetic Algorithms (GA).

Origin Genetic algorithm pseudo-code:

t=0;
Initialize $P(t)$;
Evaluate structures in $P(t)$;
Repeat
t= t+1
Select- reproduction $C(t)$ from: $P(t-1)$;
Combine and mutate structures in $C(t)$ forming $C'(t)$;
Evaluate structures in $C'(t)$;
Select-replace $P(t)$; from $C'(t)$ and $P(t-1)$;
Until (termination condition satisfied) [16].

Once the genetic operators have been applied, the local maximum fitness value is computed and then compared to the global maximum. If the highest value within a specific region is larger than the highest value across all regions, then the highest value across all regions is replaced with the highest value within that specific region, and the following iteration proceeds with the updated population. The cluster spots will be relocated based on the chromosome that exhibits the highest global value. Alternatively, the subsequent iteration proceeds with the unchanged population. This process is iterated N times. The subsequent section demonstrates that our refining approach enhances the quality of the clusters. The algorithm is as follows:

1. Declare full features or attributes of *PHIC4PHC-42* that generate as initial population $(p1)$.
2. Calculate the distance between population in each chromosome using K-NN as fitness function.

3. Choose the chromosome with the highest fitness value, then store it as global maximum (G_{max}).
   a. For i = 1 to L do
      i. Perform reproduction
      ii. Apply the crossover operator.
      iii. Perform mutation and get the new population. ($p2$).
      iv. Calculate the local maximum (L_{max}) using K-NN as fitness function.
      v. If $G_{max} < L_{max}$ then
         1). $G_{max} = L_{max}$
         2). p1 = p2;
   b. Repeat until convergent
4. Output —The last chromosome configuration, which is convergent with new smallest feature subset that obtains, G_{max} will be established. The chromosome has the optimum accuracy with classification K-NN, will be known as the best new configuration of features performance [17].

3.2 K-NN Algorithm

The K-NN algorithm is a pattern recognition algorithm that consistently achieves great performance in experimental results across multiple datasets. K-NN is a supervised learning algorithm and a significant non-parametric method [13]. The classification rules were constructed solely from the training samples, without the need for any extra data. The k-NN classification technique utilizes the k-nearest training samples to predict the category of a test sample. It assigns the test sample to the category with the highest category probability among its nearest neighbours. The procedure for applying the K-NN algorithm to categorize sample X involves the following steps:[18]:

1. Suppose there are j training categories $C_1, C_2, \ldots\ldots C_j$ and the sum of the training sample is N after feature reduction. They become m-dimension feature vectors;
2. Make sample X to be the same feature vector of the form $(X_1, X_2, \ldots\ldots X_m)$ as all training samples;
3. Calculate the similarities between all training samples and X. Taking the i^{th} sample ($d_{i1}, d_{i2}, \ldots\ldots d_{im}$) as an example, the similarity $SIM\ (X, d_i)$ is as following:

$$SIM\ (X, d_i) = \frac{\sum_{j=1}^{m} X_j, d_{ij.}}{\sqrt{\left[\sum_{j=1}^{m} X_j\right]}\sqrt{\left[\sum_{j=1}^{m} d_{ij}\right]}}$$

4. Choose k samples which are larger from N similarities of $M(X, d_i)$, (i=1, 2,…, N), and treat them as a K-NN collection of X. Then, calculate the probability of X for each category, with the following formula.

$$P(XC_j) = \sum_{d} SIM\ (X,d_i) \,.\, y\ (d_i, C_j)$$

Where $y(d_1, C_j)\ is$ a category attribute function which is satisfied

$$y(d_1, C_j) = \begin{cases} 1, d_1 \in C_j \\ 0, d_1 \notin C_j \end{cases}$$

5. Judge sample X to be the category which has the largest $P(X, C_j)$.

4 Results and Discussion

4.1 Results

4.1.1 Experiment Result

This study conducted an experiment to determine the optimal and most suitable model by selecting features with high accuracy and the most effective feature composition.

Table 2. The comparison of features and classification accuracy among the existing questionnaire and the experiment model.

Type	Classification Accuracy	Root Mean Squared Error
PHIC4PHC 42	91.7% +/-2 .46% (micro average: 91.70%)	0.244 +/- 0.03 (micro average: 0.246 +/- 0.000)
Experiment		
8 F	88.19% +/- 5.28% (micro average: 88.16%)	0.313 +/- 0.049 (micro average: 0.286 +/- 0.000)
9 F	88.94% +/- 5.54% (micro average: 87.93%)	0.295 +/- 0.53 (micro average: 0.300 +/- 0.000)
10 F	89.17% +/- 6.45% (micro average: 90.31%)	0.295 +/- 0.066 (micro average: 0.302 +/- 0.000)
11 F	88.62% +/- 4.63% (micro average: 88.60%)	0.290+/-0.063 (micro average: 0.296 +/- 0.000)
12 F	90.13% +/- 5.97% (micro average: 90.15%)	0.272+/-0.074 (micro average: 0.282 +/- 0.000)
13 F	88.72% +/- 5.69% (micro average: 88.72%)	0.295 +/- 0.058 (micro average: 0.300 +/- 0.000)
14 F	91.49% +/- 4.04% (micro average: 91.48%)	0.261 +/- 0.058 (micro average: 0.267 +/- 0.000)
15 F	90.83% +/- 5.22% (micro average: 90.82%)	0.269+/-0.066 (micro average: 0.277 +/- 0.000)
Parameter		
Optimize Selection: Population Size: 5 Maximum Number Of Generation: 30 Normalize: Weight	Cross-validation: Number of the fold: 30 Sampling type: automatic Parallel execution: Enable	

Maximal Fitness: Infinity	k-NN:
Selection Schema: Tournament	k: 1
Tournament Size: 0.25	Measure type: Mixed Measure
P Initialize: 0.5	Mixed Measure: Mixed Euclidean Distance
P Mutation: -1.0	
P Crossover: 0.5	

Table 2 demonstrates a negative correlation between the size of the characteristics and the accuracy of the categorization, as shown in both the existing questionnaire and the model. Table 2 shows that the inclusion of an additional feature in the model generally leads to an increase in accuracy, with the exception of the models including 13 and 15 features.

4.1.2 Dimensional Composition

The original edition of PHIC4PHC comprised 10 indicators encompassing 42 different features. The experiment yielded many model questionnaires, each consisting of 8-15 attributes. Every questionnaire contained a range of parameters.

Table 3. Distribution Features among result experiments

Indicator	**42**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Cognitive Proficiency									
HIS Knowledge	1-8	7,8	6,7	6,7	1,6,7,8	7	6,7	6,7,8	3,6,7,8
HIS Skills	9-11		9	9			9,11		10
Technical Proficiency									
General Computer Skills	12-21	16,19	13	13	13,16 18,19	14,16	13,15,17	12,14,16,21	12,16,21
Office Skills	22-30	30	24,25,27,28	25,27,28,30	22,30	18,19,23,30	25,27,28,30	23,24	23,24,30
Network Skill	31-32							32	32
Ethical Proficiency									
Security and Legal Knowledge	33-34					33,34		33	33
Health Information Literacy									
Access	35-36	35,36			36	35		35	
Manage	37,		37	37			37	38	

Indicator	42	8	9	10	11	12	13	14	15
Cognitive Proficiency									
	38								
Integrate	39-40	40		40		39,40	40		38
Evaluate	41-42							41	41

Table 3 outlines the indications frequently observed in the experiment's results, which include competency in the field of Health Information Systems (HIS), proficiency in computer and office skills, and integrated information literacy.

This study proposed a questionnaire called PHIC4PHC-S12Q, which consists of 12 features. The model PHIC4PHC-S12Q was selected due to its fulfillment of the primary indicator and the inclusion of all indicators in ethical proficiency, distinguishing it from other models. The accuracy of PHIC4PHC-S12Q is 90.13%, which is 1.57% lower than the original version.

4.2 Discussion

There is a growing global agreement that health systems must be equipped with digital capabilities in order to consistently enhance their performance. The issue of defining and measuring digital excellence in health care, which refers to the safe and effective utilization of digital health technologies, is a topic of great interest. [19]

Public health practitioners in the 21st century have significant obstacles, related to advancements in technology and shifts in demographics (Hernandez et al., 2003). The increasing prevalence of health IT adoption and its essential role in the efficient execution of tasks in public health care (PHC) has made public health informatics competencies (PHIC) crucial for PHC personnel. [20]

The PHIC4PHC is a pioneering questionnaire designed to evaluate the essential abilities needed by primary healthcare workers in the digital health era, including proficiency in computer usage, making ethical decisions, and health literacy. [10]

Data mining techniques have been employed within the health community to address public health issues. Public health data mining is utilized for monitoring and gathering health information from social media platforms, as well as for analyzing health behavior and managing public health services. [14, 21, 22]. This study utilized an optimal selection methodology, namely a genetic algorithm, as a version of the feature selection operator for the search strategy. Genetic algorithms (GAs) are resilient machine learning methods used to condense a vast array of variables into a smaller subset that effectively captures the most variability included in the initial data. A genetic algorithm was utilized in studies to condense the questions of the questionnaire The conventional approach employed for abbreviation was the utilization of R application. Nevertheless, the latest research have not provided any information regarding the model's performance and the accuracy of categorization based on the shorter questionnaire. This study using data mining techniques, especially feature selection with genetic algorithm, to test the correctness of an abbreviated questionnaire.[13, 14, 23].

This study proposed PHIC4PHC-12Q as the best experiment result even though the accuracy is lower compare than 14 and 15 questions. This model has a completed security and legal knowledge indicators as the focus point. Confidentiality and privacy concern persist regarding the adoption of ICT in health care organizations, especially in low to middle income nations like Indonesia.

5 Conclusions

A data mining technique employing feature selection was utilized to create a concise version of the questionnaire, which also demonstrated high accuracy in comparison to the original versions but with less questions. The utilization of both the genetic and k-NN algorithms enhanced the accuracy of feature selection for label prediction despite the inclusion of a smaller number of characteristics in the questionnaire design. This study proposed the short of PHIC4PHC questionnaire consist of 12 questions (PHIC4PHC-S12Q) as the alternative questionnaire to measure public health informatics. Future studies need to conduct the confirmatory factor analysis to compare the model construct by data mining technique.

References

Martynov, V.V., D.N. Shavaleeva, and A.A. Zaytseva. *Information Technology as the Basis for Transformation into a Digital Society and Industry 5.0*. in *2019 International Conference" Quality Management, Transport and Information Security, Information Technologies"(IT&QM&IS)*. 2019. IEEE.

Rabeh Morrar, a. Husam Arman, and S. Mousa, *The Fourth Industrial Revolution (Industry 4.0): A Social Innovation Perspective.* Technology Innovation Management Review, 2017. **7**(11).

Sachs, J.D., *Achieving universal health coverage in low-income settings.* The Lancet, 2012. **380**(9845): p. 944-947.

Rohde, J., et al., *30 years after Alma-Ata: has primary health care worked in countries?* The Lancet, 2008. **372**(9642): p. 950-961.

Rachmani, E., et al., *The implementation of an integrated e-leprosy framework in a leprosy control program at primary health care centers in Indonesia.* Int J Med Inform, 2020. **140**: p. 104155.

Ludwick, D.A. and J. Doucette, *Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries.* Int J Med Inform, 2009. **78**(1): p. 22-31.

Greenhalgh, T. and J. Russell, *Why do evaluations of eHealth programs fail? An alternative set of guiding principles.* PLoS Med, 2010. **7**(11): p. e1000360.

Heeks, R., *Health information systems: failure, success and improvisation.* Int J Med Inform, 2006. **75**(2): p. 125-37.

Afrizal, S.H., et al., *Barriers and challenges to Primary Health Care Information System (PHCIS) adoption from health management perspective: A qualitative study.* 2019. **17**: p. 100198.

Rachmani, E., et al., *Development and validation of an instrument for measuring competencies on public health informatics of primary health care worker (PHIC4PHC) in Indonesia.* Prim Health Care Res Dev, 2020. **21**: p. e22.

Rachmani, E., et al. *Si Cerdik (Information System How to Evaluate the Digital Range of Health Information Literacy).* 2022 [cited 2022; Available from: https://sicerdik.dinus.ac.id/kategori-dhlc/.

Suárez Sánchez, A., et al., *Applying the K-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders.* International Journal of Industrial Ergonomics, 2016. **52**: p. 92-99.

Sahdra, B.K., et al., *Using Genetic Algorithms in a Large Nationally Representative American Sample to Abbreviate the Multidimensional Experiential Avoidance Questionnaire.* Front Psychol, 2016. **7**: p. 189.

Rachmani, E., et al., *Developing an Indonesia's health literacy short-form survey questionnaire (HLS-EU-SQ10-IDN) using the feature selection and genetic algorithm.* Comput Methods Programs Biomed, 2019. **182**: p. 105047.

Konak, A., D.W. Coit, and A.E. Smith, *Multi-objective optimization using genetic algorithms: A tutorial.* Reliability Engineering & System Safety, 2006. **91**(9): p. 992-1007.

Yan, X., *Weighted K-nearest neighbor classification algorithm based on Genetic Algorithm.* Indonesian Journal of Electrical Engineering and Computer Science, 2013. **11**(10): p. 6173-6178.

Suguna, N. and K. Thanushkodi, *An improved k-nearest neighbor classification using genetic algorithm.* International Journal of Computer Science Issues, 2010. **7**(2): p. 18-21.

Lihua, Y., D. Qi, and G. Yanjun, *Study on KNN text categorization algorithm.* Micro Computer Information, 2006. **21**: p. 269-271.

Cresswell, K., et al., *Reconceptualising the digital maturity of health systems.* 2019. **1**(5): p. e200-e201.

William A. Yasnoff, et al., *Public Health Informatics: Improving and Transforming Public Health in the Information Age.* Journal Public Health Management and Practice, 2000. **6**(6): p. 67-75.

Rachmani, E., et al. *Are the Citizens of Semarang ready for The Citizen Health App (SatuSehat Mobile)? A Prediction Model from Decision Tree*. in *2023 International Seminar on Application for Technology of Information and Communication (iSemantic)*. 2023. IEEE.

Rachmani, E., et al. *Mining Medication behavior of the completion leprosy's multi-drug therapy in Indonesia*. in *2018 International Seminar on Application for Technology of Information and Communication*. 2018. IEEE.

Eisenbarth, H., S.O. Lilienfeld, and T. Yarkoni, *Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory-Revised (PPI-R).* Psychol Assess, 2015. **27**(1): p. 194-202.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.