Evaluation of an observer training curriculum using the Multifocal Approach to the sharing In the Shared Decision Making (MAPPIN'SDM)

- An instrument for evaluating shared decision making in patient consultations

BY: MARIA KRISTIANSEN, MK -12

Teaching supervisor:
Jürgen Kasper
Faculty of Health Sciences, University of Tromsø

Co-supervisor:
Eirik Ofstad
Faculty of Social Medicine, University of Tromsø

Master thesis: MED-3950 class of 2012

Tromsø: Program of Professional Study in Medicine, Faculty of Health Science

UiT, The Arctic University of Norway, 2016

UNIVERSITETET I TROMSØ

UiT / NORGES ARKTISKE UNIVERSITET

Preface

Shared Decision Making (SDM) is considered the gold standard regarding decision making in clinical consultations (1). Whether SDM can contribute to positive patient outcome is widely discussed and still requires more research. However, it is acknowledged that SDM may lead to positive cognitive-affective outcomes, which includes increased patient satisfaction and knowledge, and therefore less decisional conflict (2, 3). SDM is endorsed politically in Norway, and several recent white papers from The Ministry of Health Services states that SDM is a tool to be implemented in patient care (4). The white papers states that patient involvement is important, and gives the patient a better foundation for patient adherence. It may also decrease the likelihood of unnecessary treatment and contribute to the patient's perception of treatment success, based on the patients preferences (4, 5). From an ethical viewpoint, the patient's right to complicity and information is easy to defend, but this is also embedded in the law (6).

Currently there are no Norwegian instruments to evaluate how and to what degree, SDM is implemented in patient consultation. MAPPIN´SDM (Multifocal Approach to Sharing in Shared decision making) is a validated inventory comprising observer scales and questionnaires to measure the extent to which SDM realized in medical consultations. The instrument was developed based on the OPTION scale (7), and includes indicators for patient involvment and criteria for evidence-based patient information. MAPPIN is the first of its kind to combine three perspectives (patient, doctor, observer) when evaluating communication in medical consultations. Evaluations of trained observers based on the MAPPIN manual have been proven highly accurate and valid (8). A newly translated norwegian version on MAPPIN'SDM has recently been validated (9). This study is a part of the reasearch that is meant to validate the content of the recently translated Norwegian version of the instrument. As this is an observation based instrument, the quality of observation training is an vital element in this process.

The purpose of this study is to evaluate and assess the Norwegian MAPPIN'SDM observation training curriculum, and whether it can enable raters to assess SDM objectively and accurately.

First and foremost, I would like to thank my supervisors Jürgen Kasper and Eirik Ofstad for their helpful advice throughout the entire process. I would also like to thank Simone Kienlin and Signe Olsborg for completing the training course with me, and for all of our helpful discussions. Thanks to the doctors and the patients, providing records of their own consultations for the training purpose.

Tromsø 26.05.16

Maria Kristiansen

Index

Abstract

Background: In recent years, the concept of shared decision making (SDM) has gained support in the medical community. SDM has also been supported by legal documents as the best model of medical decision making. In light of this, development of instruments measuring the degree of patient involvement in clinical consultations has increased. MAPPIN'SDM is the first instrument to be validated and translated to Norwegian. While observation based evaluation is highly dependable on the competency of observers, empirical studies on ideal models for rater training is lacking. This study aims to evaluate and adjust the Norwegian rater training curriculum for MAPPIN'SDM. **Method**: In this feasibility and validation study, a standardized 4 days rater-training program was applied to a group of three trainees with no previous experience with the instrument. Ratings performed on videotaped clinical consultations relevant for SDM were used to calculate inter-rater reliability using T-coefficients (modified Cohen's kappa) and percentage agreement (%A). **Results:** Inter-rater reliabilities during the training were moderate to strong on average over the 11 items of each of three observer-scales (T_{mean}: $MAPPIN_{doctor}$ = .62, range= .41-.91; $MAPPIN_{patient}$ = .66, range= .36-1.0; $MAPPIN_{dyad}$= .59, range= .30 - .91). All trainees achieved accurate compared to the reference standard with regard to both sensitivity ($sensitivity_{mean}$: $MAPPIN_{doctor}$= 90, range= 56 - 100%; $MAPPIN_{patient}$= 83, range= 38 - 100%; $MAPPIN_{dyad}$= 92, range: 64 - 100%) and specificity ($specificity_{mean}$: $MAPPIN_{doctor}$= 83, range= 56 - 97%; $MAPPIN_{patient}$= 81, range: 33 - 100%; $MAPPIN_{dyad}$= 90, range: 81 - 100%). The results also show that the new curriculum is capable to develop high to excellent rater competency within a 4 day rater-training program. **Conclusion**: The observer training curriculum corresponding to the MAPPIN'SDM observer scales proves feasible and capable to develop high observer competencies. In addition, the study reveals a need for evaluated trainings necessary for making use of observation-based communication assessment scales in general.

Background

Shared decision-making (SDM) is the term for a communication of patients making informed medical decisions supported by concerned health personnel. The key operation is a non-coercive exchange of information based upon an agreement that it is up to the patient to evaluate and consider possible benefits against possible harms (10).

SDM is increasingly seen as best practice model of medical decision making. This is reflected in ethical guidelines (11), legal documents (6), and in many countries, Norway included, by the current course of public health policy making (4, 12, 13). A vast majority of patients would prefer SDM to traditional communication, and wish to be informed about the available options (14).

Despite clear signs of a culture change in health communication, SDM is to the best of our knowledge, yet to be completely implemented in any health system or clinical practice (15). This implies a continuous need for evaluation of means designed to facilitate patient involvement. A large number of instruments have been developed to assess consultations with regard to whether SDM is being realized, and to what extent (16-19). Some of instruments work based on structured observations (20). MAPPIN'SDM provides an inventory comprising SDM assessment scales to be administered by either the involved parties themselves (doctors and patients) or observers rating video records of the communication (18). All scales consistently use an identical set of indicators. The inventory has repeatedly shown good reliability and validity, and this presents itself in several languages (English, German, Dutch, Serbo-Croatian, Italian, Norwegian). Responding to an evaluation of a rater training (21), the MAPPIN'SDM inventory has recently been revised. The set of 15 indicators was restructured without loss of information to a shortened and presumable more distinctive solution comprising of 11 indicators (8). In appraising observer instruments, rater-training has a crucial role. As psychometric properties of

observation-based instruments are achieved by combining both measurement items and the competency to use them properly, administration of such methods requires availability of skilled observers with proven inter-rater-agreement. Rater training may contribute to increase the inter-rater-reliability and rater competency, to ensure objective and accurate observations. Due to a lack of guidance regarding efforts and methods needed to calibrate the instrument, potential users (e.g. researchers) might be reluctant in relation to choosing an observer measure. As didactic design of rater-trainings is not trivial and to some extent specific for the particular measure, evaluation of observer training methods should be considered an essential part of developing an observer measure.

Status on research

I conducted a systematic literary search using the database MEDLINE (ovid). The previous was done in order to find any evidence on curricula used to ensure rater reliability between raters using an observation-based scale to assess any quality of communication. I combined search terms of two groups with AND, within group terms with OR. The first group included synonyms of the search term “observer instrument”, the second group represented terms used for “rater training”. Due to few results, a third group comprising terms around “didactic methods” was later removed. The search revealed 27 hits. References were considered potentially relevant by this author and my supervisor JK when indicating English written empirical studies about training observers in coding communication quality. Amongst 10 full-texts selected for closer consideration I identified 4 articles as relevant (22-25).

Three studies tested inter-rater-reliability after a training in scoring the Hamilton Depression Rating Scale (HAMD). As the MAPPIN’SDM, HAMD is applied to video records of consultations during rater training. HAMD rater training is studied in three applications, but varies with regard to the setting of the training, target group and methods used. Inter-rater-reliability and validity with regard to

an expert standard are used as outcome variables. The studies demonstrate efficacy of the training both applied to novices and to participants with more pre-experience. Trainings were provided individually, in groups, and in context of an online tutorial. The most important methods were lectures, expert guidelines, example videos, and feedback and group discussions. The authors also reported on materials used, e.g. the type of videos, an introductory lecture or expert guidelines, though this was done without providing further details. By the indicated number of videos and sessions in the trainings, readers get an estimate on the required quantity of training to achieve a satisfactory agreement.

The fourth study on the HAMD and related psychiatric scales evaluated moderators of rater competency, such as extent of previous clinical experience. Results are useful to inform selection of suitable candidates for rater trainings (25).

Summarizing, there is little evidence on didactic methods used to achieve IRR (interrater reliability) /ICC (interrater correlation coefficient) with regard to both number of studies and detail of methods` description. Our search identified one series of studies referring to one observer scale providing the kind of information that we consider essential for users of observer scales in general. No studies were published comparing different methods with regard to e.g. time needed or resulting degree of agreement. This review implies a need for studies describing and testing rater training didactics of corresponding observer instruments.

This study aimed at evaluating an observer-training curriculum corresponding with the newly revised and validated version of the MAPPIN'SDM. In particular, this study focused on the feasibility of the training with regard to practical issues, time, usability and comprehensibility of the materials, learning settings and teaching communication. Moreover, this study investigated the training's capability to enable raters using the scales in a reliable manner. As both foci, feasibility of the training and inter-rater-reliability are strongly inter-related, this study aims at identifying potential need for revision of either the scale and its

indicators or the manual and corresponding working materials or the didactic methods used within the training. This study was conducted to allow for provision of reliable information on training needs to other researchers considering using the MAPPIN'SDM.

Methods and material

Design

Our study on feasibility was designed using both qualitative and quantitative methods that were applied to the first use of a new version of the rater training. Focusing on usability and comprehension, the study implied a pilot test of the corresponding materials, items and procedures. These processes resulted in a detailed review including recommendations for revision. Due to the concept of an observer based instrument consisting of a composition of both the materials (rater sheet and manual) and the observer making use of these materials, feasibility was also studied by focusing on this interaction. This implied measuring the extent to which observers during the course obtained agreement in their judging and, whether observers became capable of presenting valid judgments compared to a reference standard. Inter-rater reliability (IRR) and validity were used as criteria by which the curriculum objective was considered to be attained. Applied to another material, sustainability of IRR was tested 4 months after the training.

The project has previously been approved by the Norwegian Regional Ethical Committee, and was not deemed a subject necessary for a new application and disclosure. All patients and doctors have signed informed consents obtained by the researcher who collected the material.

Sample

The present study used a convenient sample of young scientists with a background within medicine or nursing. Amongst a bigger group of interested

individuals a group of three raters was identified complying with the criteria of availability to a given time frame of 4 working days, being interested in measuring communication quality and having basic knowledge on medical issues. Although not a criterion for participation, all participants were initially interested to join beyond training the health communication research group. The group consisted of two third year medical students, this author included, and a nurse completing a master degree on health and empowerment.

Curriculum

The revision of the inventory after and based on the testing of the previous rater training implied revision of the curriculum too (26). Firstly, the new curriculum had to deal with the restructured and shortened set of indicators. Secondly, the manual and corresponding teaching materials required adaption according to the new structure. Thirdly, as the training was conducted in Norway, all materials, including example consultation videos, were newly developed in the Norwegian language.

The present curriculum comprised of [five] didactic units, which successively were conducted within a four days intensive workshop and includes the testing of the resulting IRR:

[1] SDM education: To establish a basic understanding of the concept and an idea of how patient involvement is realized in clinical encounters, an introductory lecture on this concept was presented to all trainees. In a narrative manner, trainees were introduced to; the shared decision making story; distinctions to paternalistic communication; various approaches to training of health professionals and research projects. The presentation was enriched by providing video examples of SDM and demonstration of other decision support strategies such as the decision aid platform “mine behandlingvalg” (27) and further, the three question method (28), which is developed to support patients’ active involvement into the communication was also demonstrated. Aiming at achieving greatest possible identification with the subject, possible related research

questions and study ideas were localized on a mind map and offered to the trainees. It was known from earlier trainings that trainees with interest in related research show most stable motivation. During the discussion, the subject was positively connoted and trainees were invited to be participants in an innovative movement. The second educational sequence within the curriculum was devoted to [2] Evidences Based Patient information (EBPI) as a key element in Shared Decision Making. Trainees learned to conceive the SDM communication method as vehicle of Evidence Based Medicine (EBM), which aims at finding the scientific evidence from groups appropriate decisions for the individual patient. This understanding implies the need to involve patients by sharing the evidence in a way patient can process (29). Instead of requiring studying basic literature before starting the training, the trainees received a [3] literature workload during the first day of the training (8, 18, 29 - 32) and the MAPPIN manual.

Use of EBPI criteria for assessment of patient information was practiced using print information examples typically provided to patients by the local hospital. Moreover, trainees got insight into research on different risk figure presentation formats used to present study effects. The first two didactic units within the curriculum were set up flexibly to facilitate interactivity and adjustment to the training group. In total, these two units were dealt with during the first four hours. It should be emphasized that the first day should not be finished without providing a closer look into the measurement method. After a 10 slide PowerPoint introduction of [4] the MAPPIN'SDM approach, each of the eleven indicators for SDM were explained in detail, and examples were provided. Presentation of the criteria for the five scoring levels was taught using the MAPPIN'SDM manual. Further, the first consultation-video was watched and appraised within a moderated group discussion. A buster session of the MAPPIN-approach unit was also given the other day, and the questions that the trainees brought from their home studies were also answered.

The following three days were spent for [5] interactive observer-training to approach and to prove a satisfying level of agreement within the group and with

the expert standard. If needed, the observer training was interrupted to provide additional background information, answer questions to the rater training, or drafting several study ideas. Timeouts such as those just mentioned, were important to maintain endurance and motivation for the stereotypical coding procedure. In total, 25 videos were assessed over the course of the training, 19 of which were used to demonstrate IRR. With raising pace, the videos were administered always using the same 6 steps:

1. Briefing (structural and regarding medical issues)
2. Rating independently
3. Discourse
4. Finding consent
5. Documentation in EXCEL sheet and
6. Expert briefing

This proceeding was followed rigorously to achieve both sufficient number of ratings for calculation of IRR and stepwise consolidation of observer competency. The underpinning mechanism addressed by this proceeding is a social validation process were individual social perception is calibrated to approach common ground.

The video material was selected based on relevance and consisted of real clinical consultations between a patient and a doctor, where a decision regarding treatment or diagnosis was being questioned. The videos were recorded by medical specialists from the University Hospital of North Norway (UNN), to which this project is affiliated.

Measurement

Initially, trainees provided informed consent for both the training itself and participation in the evaluation of the curriculum. On the one hand, this implied making the rating data available for analysis of IRR and reporting these results within a scientific publication. On the other hand, participation in the evaluation of

the curriculum required contributing to identifying potential needs for revision in the manual, the teaching method or the practice sample of videos. This meant in particular, that the first conduct of the adapted rater training curriculum was continuously accompanied by a meta-communication in the training group on feasibility issues. Within each discourse session, attention was given to an analysis of reasons for misconceptions. This was done by e.g. in depth interview sequences, the observation of usage of the study materials by the moderator, or by initial utterances by the trainees. An example would be: if the wording in the manual subsequently led to individual interpretation by raters resulting in different rating scores, the phrasing in question was evaluated. Barriers towards comprehension identified during the training, were documented to be used in a following revision.

To prepare the quantitative measurement and to provide identical information to all raters involved, decision sequences had been coded a priori with regard to timeline, type of decision (diagnostic, treatment, medical domain) and the set of available options. If necessary, medical expertise was requested to affirm the given set of available options.

Within the training course, 25 decision sequences underwent observation-based analyses by the three trainees and the SDM expert who was moderating the course. All coders worked independently and were unaware of each other's ratings. Sequences were selected in random order. Single ratings were documented to allow for calculation of inter-rater-agreement, before a consensus rating was agreed upon through discourse. However, videos rated within the first two days were not used for IRR check. Rater competency was tested at two occasions. First: Data obtained within the last two days of the training course, were used to calculate inter-rater-reliability and validity. Within this test 19 decisions were coded in total. Five months after the training, another test of reliability between two finishers of the training was conducted. In the test aforementioned, a new sample of 35 medical decisions was used as a test set.

Raters used a Norwegian translated version of the MAPPIN'SDM instrument (8). Raters in the present study had to provide judgements on 11 indicators according to each of the three observation foci; doctor, patient and dyad (table 1). Each item is rated from '0' to '4' where '0' represents 'The behavior is not observed" and '4' represents 'The behavior is observed to an excellent standard" (8). The expert provided in a dichotomous format: "SDM present" [1], or "SDM absent" [0] for each item and as general judgements for each sequence. In absence of a gold standard, these judgements worked as a reference standard of SDM (33).

An overall evaluation was added by the moderator after the finish of the training and the calculation of the resulting IRR, to consider appropriateness of the timeframe, number and character of the practice sample.

Analyses

Documented issues indicating need for revision were attributed to the different components of the training, rater sheet, manual, education methods, and practice videos. Suggestions in this regard and regarding e.g. reformulation of phrases in the manual, were already collected during the training. An expert panel built of researchers with long experience with the MAPPIN'SDM inventory made final decisions on whether revision was needed

Data from rating procedures presented as 15 separate series; three for each rater using the MAPPIN'SDM (MAPPINdoctor, patient & dyad / rater 1/2/3/expert = 12), and a consensus judgement for each of three MAPPIN'SDM scales. Pairwise inter-rater reliabilities were calculated within the rater-team using EXCEL sheets on single item and on mean score level based on T coefficients (34). T represents a modified Cohen's kappa using theoretical assumptions rather than empirical data to estimate expected values (31). As observers were trained to maintain awareness also with regard to less likely events, equal distribution of expected events over the scale range was considered reasonable. T values between .40 and .50 are considered moderate, higher than .60 strong and T higher than .80 excellent (35). Moreover, percentage of agreement

(percentageA) was calculated item-wise. Mean values for T and percentageA were calculated for the rater team. IRR of the 5 months follow up test were calculated using the same proceeding with the only exception that this time only two raters from the original training group were involved.

To allow for a calculation of sensitivity and specificity, MAPPIN'SDM consent-scores were dichotomized both on item and on mean score level. Judgements lower than "2" (basis competency on the MAPPIN'SDM scale) were defined as "SDM absent", judgements "2" or higher were defined as "SDM present". The cut-off used to split the mean scores was 1.49 (37.25 of 100 respectively). This was done using SPSS version 23. Using IBM's SPSS version 23, four field tables were created and values of sensitivity and specificity of MAPPIN'SDM with regard to prediction of the reference standard were calculated on item level, and the level of general judgments of the decision sequences

Work progress

The training program was completed in June 2015. The second rating done to assess sustainability of IRR was completing in November 2016 over the course of two days. The data from the rating sheets were collected during the fall of 2015 and applied to excel sheets and SPSS by the end of the year. January through February 2016 was spent collecting literature and structuring the findings. The writing process was completed according to plan in May 2016.

Results

Descriptive results of the material

The 25 videos showed clinical consultations including at least one medical decision relevant for SDM. The medical decisions discussed in these talks were related to either; oncological [16]; gynecological [4]; urological [2]; or gastro surgical [3] problems. Lengths of consultations ranged from 5.5 to 28.25 minutes (mean 15.5min). The decision sequences analyzed in this study were sometimes

shorter than the total length and at times scattered over the whole consultations. Communication quality in terms of SDM performance was low. According to the ratings made by the expert, SDM presented to at least minimal extent in about 25% of the consultations (doctor behavior 4 /19, patient behavior 2/19, dyad 6/19).

Needs for revision identified

- **Manual and coder sheet**

For the most part, the recently translated materials were perceived comprehensible and seemed to fully transfer the original meaning to Norwegian. In addition, the materials were considered consistent with the approach and with each other and detailed to appropriate extent. During the training, a few indicators were identified as unclear with regard to explanations in the manual. In particular, definitions and examples provided to guide the coding of observed events between “0” and “4” on a Likert scale were in some places perceived as misleading or unclear by the trainees. In depth interviews revealed the very nature of the comprehension problem. E.g. the criteria affiliated to the definition of level 1 (*minimal attempt*) within indicator 3c shall amongst other rules say, that the level is observed if: “some of the frequencies are presented in consideration of the EBPI criteria“. This rule did not in sufficient detail illustrate how such a mentioning could look like, what demonstration of considering EBPI criteria could be accepted, and how many of the given frequencies needed to be presented to attain the point. By further processing the documented problems in relation to comprehension, they were classified as either translation mistakes (including lacking adjustment to cultural issues) or communication problems, which already were present in the original materials but hitherto had not been recognized. At an almost equal extent we found a need for revision in both categories.

- **Proceedings / teaching methods/ selection of participants**

Despite the high work load and endeavor required to maintain high concentration over four days, the training course was considered informative, motivating and interesting by all participants and the moderator. Waiving a preparing home study

was not observed unfavorable to start the education program with. In turn, the participants were motivated to read additional articles after the first course day. Structure and division of time were appropriately useful, but adhering to the time schedule was considered even more important. Trainees perceived the group discussions as most important to achieve the competence of a SDM coder. This included initial discussions supposed to establish an identification with the SDM concept in each of the trainees. From the moderator's point of view, previous medical knowledge as contributed by the medical students was even more important as an already existing dedication to communication issues.

- **Video material**

The 25 videos used in this study had been recently recorded and were authorized for the first conduct of the training only. This limitation was at least helpful for collection of the videos but means on the other hand disproportionate efforts as a new training would require a new training pool. Instruction of the recording doctors proved appropriate, as all consultations included SDM relevant medical decisions. All consultations were realistic with regard to both patient and doctor behavior. Length of consultations turned out partly obstructive, as the learning gain from one video is independent of its length, while training time and concentration of the trainees are the most limiting factors. Although the communication sample was quite representative with regard to extent of realized patient involvement, the rarity of appearance of many of the MAPPIN'SDM indicators lead to an increased training time.

Inter-rater-reliability

Inter-rater reliabilities during the training were moderate to strong on average over 11 items in each of three observer-scales (table 2): (T_{mean}: $MAPPIN_{doctor}$ = .62, range= .41-.91; $MAPPIN_{patient}$ = .66, range= .36-1.0; $MAPPIN_{dyad}$= .59, range= .30 - .91). On single item level, T showed low agreement (below .40) for $4/8_{patient}$, $4/8_{dyad}$, moderate agreement (.40 - .60) for $1/3b/4/5/6/8_{doctor}$, $3b/6_{patient}$, $1/3b/5/6_{dyad}$, strong or excellent (>.60) for the remaining 17 items. Percentage agreement between the three raters for each item ranged from 44% to 100%

(mean$_{PA}$ MAPPIN$_{doctor}$= 69, '$_{patient}$= 73, '$_{dyad}$= 61%). Inter-rater reliabilities five months after the training were stronger than during the training (T$_{mean}$: MAPPIN$_{doctor}$ = .77, range= .57-.93; MAPPIN$_{patient}$ = .82, range= .61-1.0; MAPPIN$_{dyad}$= .77, range= .61 - .96). On single item level, T showed moderate agreement (.40 - .60) for 3b$_{doctor}$ and strong to excellent agreement (>.60) for the remaining 32 items. Percentage agreement between the two raters for each item ranged from 66% to 100% (mean$_{PA}$ MAPPIN$_{doctor}$= 79, '$_{patient}$= 86, '$_{dyad}$= 82%).

MAPPIN indicator	**Doctor**		**Patient**		**Dyad**	
	T (%A)					
	MP1	MP2	MP1	MP2	MP1	MP2
1	.52 (61)	.82 (86)	.63 (70)	.82 (85)	.52 (61)	.82 (86)
2	.85 (88)	.93 (94)	.96 (97)	1 (100)	.86 (89)	.96 (97)
3a	.71 (77)	.82 (86)	.91 (93)	.96 (97)	.63 (70)	.75 (80)
3b	.43 (54)	.57 (66)	.47 (58)	.79 (83)	.47 (58)	.64 (71)
3c	.72 (77)	.86 (89)	.83 (86)	.89 (91)	.78 (83)	.86 (89)
4	.45 (56)	.61 (69)	.30 (44)	.64 (71)	.30 (44)	.61 (69)
5	.52 (61)	.86 (89)	.65 (72)	.79 (83)	.50 (60)	.89 (91)
6	.54 (63)	.61 (69)	.45 (56)	.75 (80)	.47 (58)	.61 (69)
7	.91 (93)	.89 (91)	1.0 (100)	.96 (97)	.91 (93)	.89 (91)
8	.41 (53)	.71 (71)	.36 (49)	.61 (68)	.39 (51)	.71 (77)
9	.74 (79)	.75 (80)	.67 (74)	.82 (86)	.63 (70)	.75 (80)
mean	**.62 (69)**	**.77 (79)**	**.66 (73)**	**.82 (86)**	**.59 (61)**	**.77 (82)**

Table 2. *Pairwise inter-rater-reliabilities (IRR) on MAPPIN'SDM indicator and total score level for each of three observer scales. IRRs are calculated as T =modified Cohen's kappa and percentage agreements (%$_A$) on two occasions; MP1 - measurement obtained during the rater training and MP2 = measurements obtained 5 months after the training.*

Criterion validity

During the training, the three trainees achieved accurate MAPPIN'SDM results according to the reference standard (table 3). This applies to both sensitivity on average over three raters (sensitivity$_{mean}$: MAPPIN$_{doctor}$= 90, range= 56 - 100%; MAPPIN$_{patient}$= 83, range= 38 - 100%; MAPPIN$_{dyad}$= 92, range: 64 - 100%) and specificity (specificity$_{mean}$: MAPPIN$_{doctor}$= 83, range= 56 - 97%; MAPPIN$_{patient}$= 81, range: 33 - 100%; MAPPIN$_{dyad}$= 90, range: 81 - 100%).

	doctor		patient		dyad	
Indicator	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
1	83	89	75	67	91	92
2	100	94	100	100	100	94
3a	94	93	100	100	94	93
3b	73	80	38	74	83	93
3c	100	87	100	98	100	88
4	100	56	93	33	100	100
5	92	82	71	82	92	81
6	98	89	78	58	100	83
7	100	97	100	95	100	91
8	89	83	96	88	84	91
9	56	67	58	100	64	88
mean	**90**	**83**	**83**	**81**	**92**	**90**

Table 3. *Interrater validity measured by sensitivity and specificity for each MAPPIN'SDM indicator and total score level for each of three observer scales.*

Discussion

The present study adds to existing evidence on the theory, concept, validity and reliability of the Multifocal Approach to the Sharing in Shared Decision Making inventory to assess patient involvement in medical consultations (8, 18, 30-32). This study has provided a detailed description and validation data of a program used to develop rater competency in using the MAPPIN'SDM observer scales. As such, this study might represent the missing link enabling researchers with a need for a high quality measure of SDM making use of the given evidence. Information provided in this study is essential to both elaborate decisions on measurement methods and to use the MAPPIN'SDM in a meaningful manner.

This study explored feasibility and effects of group training curriculum for young researchers in using the MAPPIN'SDM observer scales. Our results show, that the new curriculum is capable to develop high to excellent rater competency within a four days course. This summary takes into account strong rater agreement, convincing validity with regard to a reference standard and proven sustainability of the effects. In a follow up test five month after the training, the rater competence was fully retrievable applied to another set of videos.

From a user's point of view, investment of two days for testing IRR (rating of 19 of 25 videos) might be considered disproportionate in a four days curriculum. A smaller number of tests, however, would be insufficient from a statistical point of view, and reliability check is an essential part of the training. Two arguments might nevertheless justify these efforts. As the training continues during testing the 19 videos, we assume rater training skills to further cumulate in this part of the curriculum. This would imply that the gained results cannot be seen as the result of the first two days training. Moreover, based on knowledge of the learning curve of raters in this curriculum, as shown in this study, users of the MAPPIN'SDM can consider initiation of study data analysis already at day three of the training. The possibility of using the second half of the rater training simultaneously, for a second purpose without paying reliability, might present another perspective on the high efforts of time required. Due to the potential to optimize the practice material, in particular by providing videos showing shorter consultations with higher levels of SDM performance, we assume the training time could be shortened in future conducts.

One might argue that sensitivity and specificity may be overestimated in this study as the reference judgment for our calculation has been delivered by JK, who has authored MAPPIN'SDM. As a consequence, both judgments refer to the same definition of SDM. Since the standard was determined independently and not compared to measures built upon other concepts, this proceeding might just have reduced error variance due to diverging concepts.

One might also argue that the competency obtained by the trainees are not transmissible to another material, as the video sample used for this study was not random or representative. However, it was important that the consultations were relevant examples of decision making in clinical consultations. To achieve the competency necessary to assess SDM and relevant clinical communication skills, the selection process was considered as a requisite. It is hardly imaginable that the competency to do so should not be valid when applied to a randomized selection.

The test of the new curriculum was caused by the recent revision of the inventory, within which the set of indicators underpinning the MAPPIN'SDM was restructured and shortened from 15 to 11. In a rater training of the former version (15 indicators) five training days not enough to achieve sufficient agreement (20). The revision of the indicator set was based on in depth analysis of observations within the former training. As the present training led in shorter time to stronger rater agreement, this study can be considered an indirect prove of the advantages in the MAPPIN'SDM$_{11}$ compared to the MAPPIN'SDM$_{15}$ inventory.

As the systematic literature review indicates, developers of observer scales are not used to providing research based guidance to use these instruments (21-24). In consequence, the withholding of knowledge might imply use of such instruments limited to the developers themselves. Another, but perhaps even more problematic scenario is that rater scales might be used inappropriately, due to a lack of knowledge on how to achieve measurement quality. In this regard, our study might be important both to demonstrate a type of knowledge that is largely ignored and to serve as a particular model to other observation-based instruments. Calibration of observation-based instruments to make them reliable and valid will always require additional efforts. In respect of the potentially higher data quality of observation-based compared to subjective data, these efforts might nevertheless be reasonably invested.

Conclusion

The observer training curriculum corresponding to the MAPPIN'SDM observer scales proves feasible and capable to develop high observer competencies. The study also indicated need for minor revision of the materials. The study informs researchers' decisions for or against the MAPPIN'SDM inventory and guides users to develop an effective training. Moreover, the study indicates a big and neglected need for evaluated trainings necessary for making use of observation-based communication assessment scales in general.

Table 1: MAPPIN´SDM$_{observer}$	
Indicator 1: Defining problem	
MAPPIN$_{doctor}$	The **C** draws attention to an identified problem as one that requires a decision-making process.
MAPPIN$_{patient}$	The **P** draws attention to a concrete problem as one that requires a decision-making process.
MAPPIN$_{dyad}$	**C&P** agree on a concrete problem as one that requires a decision-making process.
Indicator 2: Key message	
MAPPIN$_{doctor}$	The **C** states that there is more than one way to deal with the identified problem.
MAPPIN$_{patient}$	The **p** indicates that there is more than one way to deal with the concrete problem.
MAPPIN$_{dyad}$	**C&P** discuss that there is more than one way to deal with the concrete problem.
Indicator 3a: Options (quality of the structure)	
MAPPIN$_{doctor}$	The **C** structures the discussion of the options in a way that is easy to understand and to remember.
MAPPIN$_{patient}$	The **P** structures the discussion of the options in a way that is easy to understand and to remember.
MAPPIN$_{dyad}$	**C&P** structure the discussion of the options in a way that is easy to understand and to remember.
Indicator 3b: Options (quality of the content)	
MAPPIN$_{doctor}$	The **C** explains to the patient the pros & cons of the different options (if applicable, these include the pros & cons of 'doing nothing').
MAPPIN$_{patient}$	The **P** discusses the pros & cons of the different options.
MAPPIN$_{dyad}$	**C&P** weigh up the pros & cons of the different options.
Indicator 3c: Options (information quality)	
MAPPIN$_{doctor}$	The **C** complies with the criteria of evidence based patient information (presentation, sources, level of evidence).
MAPPIN$_{patient}$	The **P** contributes to achieving compliance with the criteria of evidence based **P** information.
MAPPIN$_{dyad}$	**C&p** consider the criteria of evidence based **P** information.
Indicator 4: Expectations & worries	
MAPPIN$_{doctor}$	The **C** explores the patient's expectations (ideas) and concerns (fears) about how to manage the concrete problem.
MAPPIN$_{patient}$	The **P** describes his/her expectations (ideas) and concerns (fears) about how to manage the concrete problem.
MAPPIN$_{dyad}$	**C&P** discuss the **P's** expectations (ideas) and concerns (fears) about how to manage the concrete problem.
Indicator 5: Indicate decision	
MAPPIN$_{doctor}$	The **C** opens the decision stage leading to the selection of an option (If applicable, deferment is a possible decision).
MAPPIN$_{patient}$	The **P** opens the decision stage leading to the selection of an option.
MAPPIN$_{dyad}$	**C&P** open the decision stage leading to the selection of an option.
Indicator 6: Follow up arrangements	
MAPPIN$_{doctor}$	The **C** makes arrangements with the **P** concerning how to proceed (e.g. steps for implementing the decision, review of decision or of deferment).
MAPPIN$_{patient}$	The **P** contributes towards the arrangements for how to proceed.
MAPPIN$_{dyad}$	**C&P** discuss plans for how to proceed.
Indicator 7: Negotiation of communication approach	
MAPPIN$_{doctor}$	The **C** ascertains the **P's** preferred approach to exchanging information (setting, media, time frame).
MAPPIN$_{patient}$	The **P** participates in deciding on the preferred approach to exchanging information.
MAPPIN$_{dyad}$	**C&P** choose an approach to exchanging information.
Indicator 8: Evaluation of patient's understanding	
MAPPIN$_{doctor}$	The **c** checks that the **P** has understood the information.
MAPPIN$_{patient}$	The **P** clarifies how he understood the information given by the **c**.
MAPPIN$_{dyad}$	**C&P** clarify whether the **P** understood the information given by the **c** correctly.
Indicator 9: Evaluation of doctor's understanding	
MAPPIN$_{doctor}$	The **C** makes sure that he has understood the **P's** viewpoint correctly.
MAPPIN$_{patient}$	The **P** makes sure that the **c** understands his viewpoint.
MAPPIN$_{dyad}$	**C&P** clarify whether the **c** has understood the **P's** viewpoint correctly.

Table 1. *The table shows the MAPPIN'SDM observer sheet. C=clinician, P= patient. Details presented in brackets added to the first (MAPPIN$_{doctor}$) of a group of three indicators also apply to the corresponding MAPPIN$_{patient}$ and '$_{dyad}$ indicators.*

References

1. Hauser K, Koerfer A, Kuhr K, Albus C, Herzig S, Matthes J. Outcome-Relevant Effects of Shared Decision Making: A Systematic Review. Deutsches Ärzteblatt International. 2015;112(40):665-671. doi:10.3238/arztebl.2015.0665.
2. Shay LA, Lafata JE. Where is the evidence? A systematic review of shared decision making and patient outcomes. Medical decision making : an international journal of the Society for Medical Decision Making. 2015;35(1):114-131. doi:10.1177/0272989X14551638.
3. Clayman ML, Bylund CL, Chewning B, Makoul G. The Impact of Patient Participation in Health Decisions Within Medical Encounters: A Systematic Review. Med Decis Making. 2016 May;36(4):427-52. doi: 10.1177/0272989X15613530.
4. Helse- og omsorgsdepartementet. Nasjonal helse- og sykehusplan (2016–2019). 2015. [Cited on 3/3/16]. Retrieved from: https://www.regjeringen.no/contentassets/7b6ad7e0ef1a403d97958bcb34478609/no/pdfs/stm201520160011000dddpdfs.pdf
5. Pasient- og brukerrettighetsloven. (1999). Lov om pasient- og brukerrettigheter LOV-1999-07-02-63 § 3.
6. Stacey D, Bennett CL, Barry MJ, Col NF, Eden KB, Holmes-Rovner M, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev. 2011 Oct 5;(10):CD001431. doi: 10.1002/14651858.CD001431.pub3.
7. Nicolai J, Moshagen M, Eich W, Bieber C. The OPTION scale for the assessment of shared decision making (SDM): methodological issues. Zeitschrift fuer Evidenz, Fortbildung und Qualitaet im Gesundheitswesen. 2012;106(4):264-71. doi: 10.1016/j.zefq.2012.03.002.
8. Kasper J, Hoffmann F, Heesen C, Köpke S, Geiger F. Completing the third person's perspective on patients' involvement in medical decision-making: approaching the full picture. Zeitschrift fuer Evidenz, Fortbildung und

Qualitaet im Gesundheitswesen. 2012;106(4):275-83. doi: 10.1016/j.zefq.2012.04.005.

9. Kienlin S, Kristiansen M, Ofstad E, Liethmann K, Geiger F, Kasper J. Validation of the Norwegian version of MAPPIN´SDM, an observation-based instrument to measure shared decision making in clinical encounters. 2016. [In submission at Patient Education Counseling]
10. Elwyn G, Edwards A, Kinnersley P, Grol R. Shared decision making and the concept of equipoise: the competences of involving patients in healthcare choices. The British Journal of General Practice. 2000;50(460):892-899.
11. General Medical Council Good Medical Practice. (2013). [Cited on 6/3/16]. Retrieved from http://www.gmc-uk.org/guidance/good_medical_practice/partnerships.asp
12. Helse- og omsorgsdepartementet. Kvalitet og pasientsikkerhet 2014. 2015. [Cited on 3/3/16]. Retrieved from: https://www.regjeringen.no/contentassets/5cd218ed18a943198ca926ec1f737855/no/pdfs/stm201520160012000dddpdfs.pdf
13. Helse- og omsorgsdepartementet. Legemiddelmeldingen — Riktig bruk – bedre helse. (2015). [Cited on 3/3/16]. Retrieved from: https://www.regjeringen.no/contentassets/1e17b19947224def82e509ca5f346357/no/pdfs/stm201420150028000dddpdfs.pdf
14. Chewning B, Bylund C, Shah B, Arora NK, Gueguen JA, Makoul G. Patient preferences for shared decisions: A systematic review. Patient education and counseling. 2012;86(1):9-18. doi:10.1016/j.pec.2011.02.004.
15. Couët N, Desroches S, Robitaille H, Vaillancourt H, Leblanc A, Turcotte S et al. Assessments of the extent to which health-care providers involve patients in decision making: a systematic review of studies using the OPTION instrument. Health Expectations. 2015 Aug;18(4):542-61. doi: 10.1111/hex.12054.
16. Barr PJ, O'Malley AJ, Tsulukidze M, Gionfriddo MR, Montori V, Elwyn G. The psychometric properties of Observer OPTION5, an observer measure of

shared decision making. Patient Education and Counseling. 2015 Aug;98(8):970-6. doi: 10.1016/j.pec.2015.04.010.

17. Clayman ML, Makoul G, Harper MM, Koby DG, Williams AR. Development of a Shared Decision Making coding system for analysis of patient-healthcare provider encounters. Patient education and counseling. 2012 Sep;88(3):367-72.doi: 10.1016/j.pec.2012.06.011.
18. Kasper J, Hoffmann F, Heesen C, Köpke S, Geiger F. MAPPIN'SDM – The Multifocal Approach to Sharing in Shared Decision Making. Reindl M, ed. PLoS ONE. 2012;7(4):e34849. doi:10.1371/journal.pone.0034849.
19. Scholl I, Kriston L, Dirmaier J, Buchholz A, Härter M. Development and psychometric properties of the Shared Decision Making Questionnaire – physician version (SDM-Q-Doc). Patient Education and Counseling. 2012 Aug;88(2):284-90. doi: 10.1016/j.pec.2012.03.005.
20. Scholl I, Loon MK-v, Sepucha K, Elwyn G, Légaré F, Härter M, et al. Measurement of shared decision making – a review of instruments. Zeitschrift fuer Evidenz, Fortbildung und Qualitaet im Gesundheitswesen. 2011;105(4):313-24.
21. Kasper J, Liethmann K, Goetze I, Geiger F. Evaluation of an observer training curriculum using the multifocal approach to the sharing in the shared decision making (MAPPIN'SDM). Oral presentation at the 7th International Shared Decision Making Conference. June 16th-19th, 2013, Lima, Peru.
22. Müller M.J., Dragicevic A. Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices. Journal of Affective Disorders. 2003 Oct;77(1):65-9.
23. Wagner S, Baskaya Ö, Lieb K, Tadic A. Standardized rater training for the Hamilton Depression Scale (HAMD17) and the Inventory of Depressive Symptoms (IDS30CR). Psychopathology. 2011;44:68–70. doi: 10.1159/000318162.
24. Rosen J, Mulsant BH, Marino P, Groening C, Young RC, Fox D. Web-based training and interrater reliability testing for scoring the Hamilton Depression

Rating Scale. Psychiatry research. 2008;161(1):126-130. doi:10.1016/j.psychres.2008.03.001.

25. Targum SD. Evaluating rater competency for CNS clinical trials. Journal of J Clin Psychopharmacol. 2006 Jun;26(3):308-10.
26. Goetze I. Trainings zur Handhabung des MAPPIN'SDM-Beobachterinstruments für geteilte Entscheidungsfindung zwischen Arzt und Patient. Kiel: Christian-Albrechts-Universität; 2013.
27. Universitetssykehuset i Nord-Norge (UNN). Minebehandlingsvalg 2015. Updated 23.11.2015. [Cited on 5/3/16]. Retrieved from: https://minebehandlingsvalg.no.
28. Shepherd HL, Barratt A, Jones A, Bateson D, Carey K, Trevena LJ, et al. The Ask Share Know Project: a research translation study of three consumer questions to enhance treatment decision making. Available from: http://www.askshareknow.com.au/
29. Bunge M, Mühlhauser I, Steckelberg A. What constitutes evidence-based patient information? Overview of discussed criteria. Patient Education and Counseling. 2010 Mar;78(3):316-28. doi: 10.1016/j.pec.2009.10.029.
30. Kasper J, Légaré F, Scheibler F, Geiger F. Turning signals into meaning--'shared decision making' meets communication theory. Health Expect. 2012 Mar;15(1):3-11. doi: 10.1111/j.1369-7625.2011.00657.x.
31. Kasper J, Heesen C, Köpke S, Fulcher G, Geiger F. Patients' and Observers' Perceptions of Involvement Differ. Validation Study on Inter-Relating Measures for Shared Decision Making. PLoS One. 2011;6(10):e26255. doi: 10.1371/journal.pone.0026255.
32. Geiger F, Kasper J. Of blind men and elephants: suggesting SDM-MASS as a compound measure for shared decision making integrating patient, physician and observer views. Z Evid Fortbild Qual Gesundhwes. 2012;106(4):284-9. doi: 10.1016/j.zefq.2012.03.020.
33. Fletcher I, Mazzi M, Nuebling M. When coders are reliable: The application of three measures to assess inter-rater reliability/agreement with doctor–patient

communication data coded with the VR-CoDES. Patient Education and Counseling. 2011 Mar;82(3):341-5. doi: 10.1016/j.pec.2011.01.004.
34. Wirtz M, Caspar F. Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Göttingen, Hogrefe, 2002.
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74.

Reference overview of central articles

Reference no. 3: “The Impact of Patient Participation in Health Decisions Within Medical Encounters: A Systematic Review” (Medical Decision Making 2016).
Authors: Marla L. Clayman, Carma L. Bylund, Betty Chewning, Gregory Makoul

Design: systematic review

Aim of the study: to evaluate whether patient involvement in clinical consultations affect patient outcomes

Method: a systematic search was conducted in PubMed to evaluate empirical evidence in doctor-patient encounters with publication date to February 28 2015.

Results: 116 articles remained after screening the 9757 that resulted from the primary search, 11 of which were randomized controlled studies. Both measurement of patient participation in medical decisions and outcome varied within the studies. Outcomes measure could be divided into 4 categories: psychosocial (e.g. decisional conflict), behavioral (e.g. adherence), practice related (e.g. encounter length) and biomedical (e.g. clinical asthma status). Most of the studies in the non-RCT group, and nearly half of the randomized control trials showed positive correlation between patient participation and at least one positive outcome. The measured outcome improved were largely psychosocial. One study also reported improvement in behavioral outcome. A negative effect was found in 5 % of the non RCTs.

Critical evaluation: the study investigates an important question and has many strengths. The review describes the search performed in detail, terms included and PICO categories, exclusion and inclusion criteria. All the authors and several assistants conducted the screening process, and at least one author and two assistants studied each article. The review may be limited by the use of only one database, although PubMed does include Medline is considered a wide database.

Reference no. 18: "MAPPIN'SDM - the multifocal approach to sharing in shared decision making" (PLoS ONE 2012)

Authors: Jürgen Kasper, Frauke Hoffmann, Christoph Heesen, Sascha Köpke, Friedemann Geiger

Design: validation study

Aim of the study: to compare relevant perspectives on patient involvement using a newly developed instrument (MAPPIN'SDM).

Method: the authors designed an instrument which included a doctor-patient-questionnaire and an observer instrument to evaluate the efficacy of shared decision making in clinical consultations. The study emphasized the importance of a bilateral approach to evaluate SDM in clinical consultations, where all perspectives (observer, doctor and patient) are combined, according to the essential understanding of SDM. The inventory was applied to 40 consultations from ten different physicians from different medial fields. Pearson correlation coefficients were used to calculate convergent validities.

Results: The results proved highly reliable. The results showed no correlation between observer judgement and doctor and patient judgement, however patient and doctor judgement were moderately related. The authors concluded, a single perspective is too limited to reach any conclusion on whether SDM was present or not.

Critical evaluation: the new instrument is described in very detail with regard to theoretical background, development process and structure of the pool of indicators. It also provides a comprehensive coder manual. The approach to measurement of patient involvement is reasonably justified, consequently and extremely systematic. The validation of the new measure is based on a sample of just 40 consultations form 10 physicians. This might challenge the certainty of the empirical data. Moreover, the validity of observational data may be limited by the presence of cameras in the consultation and hence the doctor and patient behavior. The study also points out that patient consultations were selected by the doctor included in the study. Even though they were given inclusion criteria, the selection may still be biased. This is true to most observation based method but should still be takes into consideration.

Reference no. 20: "Measurement of shared decision making - a review of instruments" (Z Evid Fortbild Qual Gesundheitswes 2011)

Authors: Isabelle Scholl, Marije Koelewijn-van Loon, Karen Sepucha, Glyn Elwyn, France Légaré, Martin Härter, Jörg Dirmaier (Z Evid Fortbild Qual Gesundhwes, 2011)

Design: systematic review

Aim of the study: as shared decision making (SDM) is being implemented in many countries, numerous instruments to measure SDM have been developed. This study aimed to systematically search the literature for published and unpublished instruments.

Method: In addition to an electronic literature search in PubMed and the Web of Science database, the authors contacted key authors in the SDM field, and also hand searched relevant journals.

Results: the authors found 28 scales used to measure shared decision making, nine of which were still in the publishing process. A dyadic approach which combines both the patients' and the physicians' perspective seems to be

trending. They concluded that while the extent of validation differed, most scales had high reliability. They emphasize the need for additional psychometric testing.

Critical evaluation: even though the authors describe a systematic search for all instruments measuring SDM, not all instruments in development were included, e.g. MAPPIN (Multifocal Approach to Sharing in Shared Decision Making). As the authors also point out, the literary search was performed using only two databases which may have led to missed publications. According to the authors the entire screening process was only completed by one individual, which may be considered a lack in systematic screening. Another limitation of the review might be its arbitrary focus on any instruments associated with SDM. The resulting pool of instruments included many instruments, which are not supposed to measure SDM but a condition, which is in any regard related to SDM. Due to a lacking idea of the SDM concept, the findings are quite unspecific and difficult to use for researchers in the field. While the study report their inclusion of unpublished scales as a strength, I question this assertion as MAPPIN'SDM was not included when it was in development.

Reference no. 22 : "Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices" (Journal of Affective Disorders, 2003)

Authors: Matthias J. Müller, Aleksandra Dragicevic

Design: validation study

Aim of the study: to implement a standardized rater-training program for the Hamilton Depression Rating Scale (HAMD). Can novice raters can accomplish acceptable inter rater reliability in three training sessions?

Method: 21 participants with various background (research students, psychologists, psychiatric residents and pharmacologists) was selected to participate in program. A standardized training program was used to train the novice raters in the use of HAMD. The program included a introductory lecture

and practical rater training using videos interviews of psychiatric patients with depressive disorders. The rating was done individually by trainees and then discussed in the group. The videos had prior to the training session been rated by expert raters. They compared trainee ratings with expert ratings for single items and the total score and calculated inter rater reliability using ICC (interrater correlation coefficient).

Results: three session were completed in total. The study concluded that inter rater reliability increased during the course of the study to a satisfactory value by the third session.

Critical evaluation: this study has a clear purpose and their method seem expedient to answer their research question. However, the study material only encompasses three videotaped interviews and is hence restricted. It is questionable whether they would be able to achieve the same results with a larger material.

Reference no. 29: "What constitutes evidence-based patient information? Overview of discussed criteria" (Patient Education Counseling 2012)

Authors: Martina Bunge, Ingrid Mühlhauser, Anke Steckelberg

Design: systematic review

Aim of the study: Evidence-based patient information (EBPI) is considered an essential part of doctor patient consultations. As EBPI is required for the patient to make an informed choice, this study aimed to review the criteria for EBPI and assemble the evidence for the criteria identified to date. It also aims to provide support for developers of EBPI.

Method: 5 databases were search to assemble the criteria for EBPI. A following search was done to find the evidence for each criterion. The authors pooled the existing categories into 13 different categories. They evaluated several studies within each category and summarized the conclusion for each category based on the review of the studies within a category.

Results: 3 systematic reviews, 24 randomized-controlled studies and 1 non-systematic review were included. The authors found the evidence to be diverse. Some of the EBPI criteria were supported by good evidence, while others resulted from ethical guidelines in clinical practice. The results include but is not limited to the following bullet points:

- The use of symbols is superior to numbers when representing the strength of a recommendation
- No study was found which evaluated the importance of patient-oriented outcome in EBPI
- Numerical presentation of risk of side effect is preferred, as patients tend to overestimate the risk when presented verbally.
- There is some evidence that picture presentation can increase level of understanding
- Cultural aspects should be considered when developing EBPI
- Plain language is recommended
- There is little agreement on which method of passing on information that will provide the greatest lever of understanding

Critical evaluation: this study has many strengths. Titles and abstracts were screened by two investigators, and they also screened the reference list. Two authors also assessed the quality and analyzed all selected papers. The included studies limited to randomized controlled trials and reviews, and all studies were screened for the risk of bias. Excluded studies were explicitly presented in a separate table. The exclusion of qualitative studies may have missed formats that showed positive results. The authors also point out that the quality varied among selected studies and may not represent an acceptable guideline.