№3 Том4 2016

Фармакоэкономика
теория и практика

Pharmacoeconomics
theory and practice

№3 Volume4 2016

- МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ ПРОВЕДЕНИЯ ОЦЕНКИ ДОСТОВЕРНОСТИ НАУЧНЫХ ДАННЫХ С ПОМОЩЬЮ СИСТЕМЫ КЛАССИФИКАЦИИ, ОЦЕНКИ, РАЗРАБОТКИ И ЭКСПЕРТИЗЫ РЕКОМЕНДАЦИЙ GRADE
- РЕЗУЛЬТАТЫ РОССИЙСКИХ ФАРМАКОЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ

www.pharmacoeconom.com

METHODOLOGICAL BASIS OF ASSESSMENT OF QUALITY OF SCIENTIFIC EVIDENCE USING THE GRADING OF RECOMMENDATIONS ASSESSMENT, DEVELOPMENT AND EVALUATION APPROACH

Ugrekhelidze D.T., Yagudina R.I.

I.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation

Abstract: *Grading of Recommendations Assessment, Development and Evaluation approach is a method of assessing the certainty in evidence and it is used in systematic reviews, health technology assessment and clinical guidelines development. In this article authors cover the process of making this assessment, the criteria of grading the evidence, determining the quality and strength of evidence.*

Key words: *GRADE, evidence-based medicine, clinical guidelines, systematic reviews, health technology assessment*

Relevance

Decision-making in healthcare should be based on reliable scientific data on the advantages and disadvantages of the considered medical technologies with a high degree of evidence. The decision makers should rely not only on the obtained positive or negative outcomes of expected use of health technology, but also on the degree of reliability of these data. Often expert groups and organizations that develop clinical practice guidelines, incorrectly interpret the available evidence base for health technologies due to the lack of systematic and unbiased system of assessment. In this regard, special attention is paid to methods that allow highly accurately and transparently evaluate the quality of scientific data. Currently there are many tools that determine the reliability of scientific data, but the most widely used technique is the one developed by a group of foreign scientists specializing in evidence-based medicine from McMaster University, Harvard University, Cochrane centres of Norway and Germany. In their works, originating in 2000, scientists drew attention to the issues of assessing the reliability of scientific information. [1] As a result, Grading of Recommendations Assessment, Development and Evaluation approach (GRADE) was developed. GRADE used to assess the validity of systematic reviews and clinical guidelines, this system is applicable to study a wide range of clinical issues, including diagnosis, screening, prevention, therapy, rehabilitation. Currently, the GRADE system used by more than 100 international and national organizations in the field of health organization, clinical pharmacology and evidence-based medicine, e.g the Cochrane community, World Health Organization, the UK National Institute for Health and Care Excellence (NICE) (table 1). [1]

Table 1. List of several organizations, using GRADE approach

- World Health Organization
- European Commission
- Cochrane community
- National Institute for Health and Care Excellence (NICE)
- Scottish Intercollegiate Guidelines Network (SIGN)
- Kidney Disease: Improving Global Outcomes (KDIGO)
- British Medical Journal
- Norwegian Institute of Public Health (Kunnskapssenteret)
- Regional Health Authority of Emilia-Romagna, Italy
- Canadian Agency for Drugs and Technologies in Health (CADTH)
- World Allergy Organization
- Agency for Healthcare Research and Quality (AHRQ)
- European Society of Thoracic Surgeons
- The National Board of Health and Welfare (Socialstyrelsen)
- European Monitoring Centre for Drugs and Drug Addiction
- Finnish Office for Health Technology Assessment
- Belgian Healthcare Knowledge Centre
- British Society of Gastroenterology

The purpose of the GRADE assessment is obtaining affordable and comprehensive information on the effectiveness of health technologies analyzing critically important outcomes for healthcare decision-making, getting data on quality and strength of evidence (i.e. the degree of confidence of the expert that the assessment of the effect performed correctly). GRADE provides a transparent and structured process of evaluation of systematic reviews and clinical guidelines, and can also be a help in the development of currently developing recommendations. During the evaluation of the quality of evidence the experts define the questions that should be answered during the analysis, choose the critically important clinical outcomes, as for patient as for healthcare specialist. The quality of the evidence is assessed taking into account possible systematic errors, inconsistency of results from different studies on one medical technology, the indirectness, imprecision in the determination of the effect size. The strength of recommendation is assessed in GRADE approach as “high” or “low” (also used the term “mandatory” and “conditional”

Ugrekhelidze D.T. - тел.: + 79162987031, e-mail: dtu1992@gmail.com

depending on the quality of scientific data and the ratio of desirable and undesirable effects from the use of the evaluated technology. [1] Figure 1 presents a schematic description of the GRADE evaluation process.

taxpayer or representatives of the insurance company). For determination of outcomes classification, it is recommended to use a nine-point scale in which clinical outcomes with the assessment 7-9 are defined as extremely

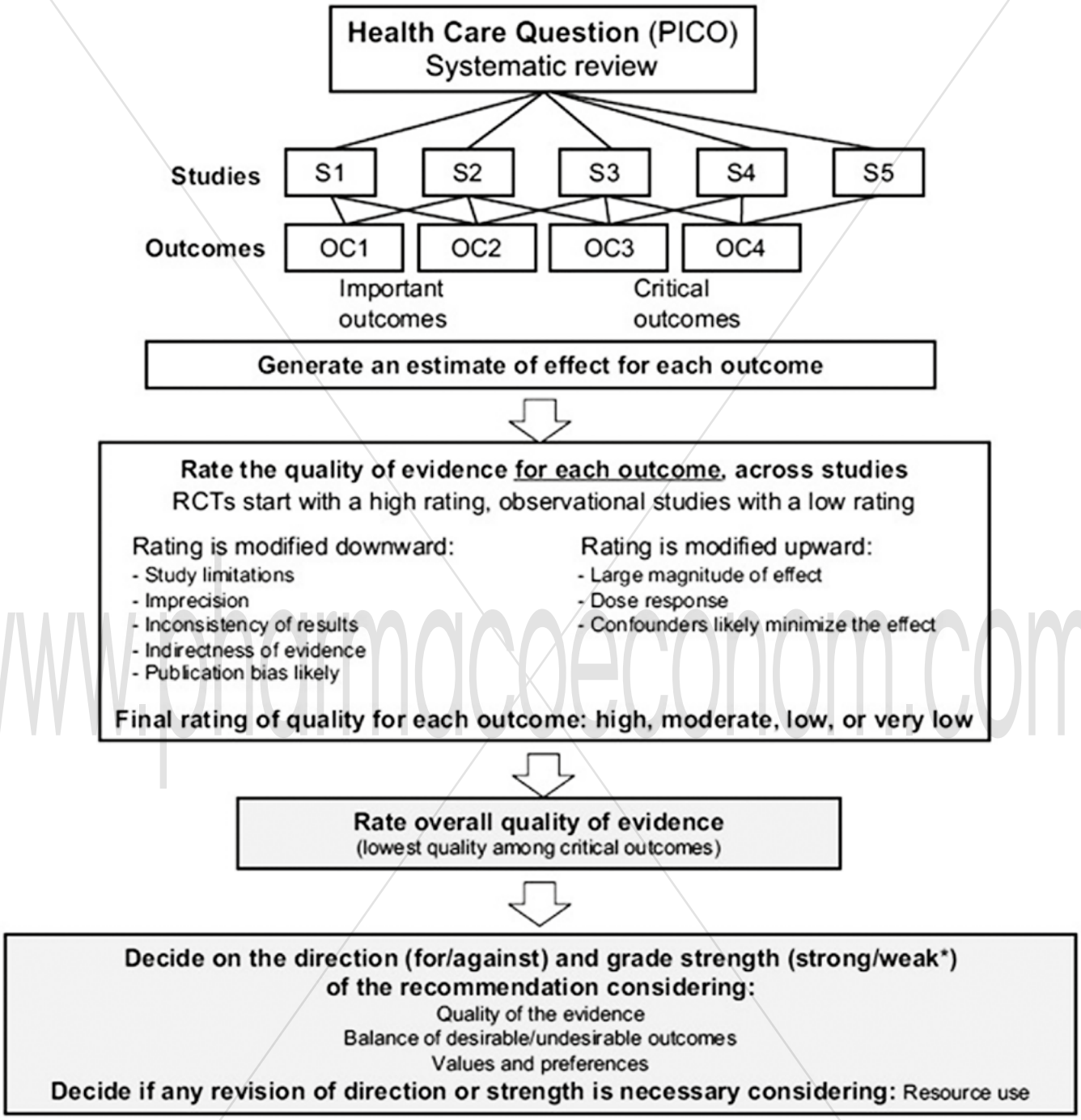

Figure 1. GRADE evaluation process [taken from Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables.]

The first step in GRADE approach is the formulation of questions using the PICO method (patient/intervention/comparator/outcome), which is determined by the population under study patients, evaluated health technology and therapy comparison of clinical outcomes. This method of identifying the most important issues in this medical problem is widely used for developing clinical guidelines and in the preparation of systematic reviews. In the preparation of clinical guidelines is also necessary to identify the most critical outcomes. The degree of importance of outcomes may vary depending on the respondent groups (patients, physicians, or health care managers), so it is important to correctly identify the key population to obtain the relevant data. The most preferable option is to incorporate the views of different points of view (e.g., opinion of a certain group of patients, the point of view of the important; a rating of 4-6 as important and with a rating of 1-3 as of limited importance) (Fig.2). Critically important and important clinical outcomes should be assessed in clinical guidelines. The use of surrogate clinical points is undesirable. They are applied only when the connection of outcomes with the state of patients. Only the most important for patient outcomes (quality of life, survival, pain) are used during evaluation of systematic reviews. [3]

The next step is to assess the quality of evidence for the critically important clinical outcomes. In the context of GRADE approach of the systematic review, the term «quality of evidence» means the degree of confidence of the expert that the assessment of the effect held true. In the context of developing clinical guidelines "the quality of evidence" means the degree of confidence

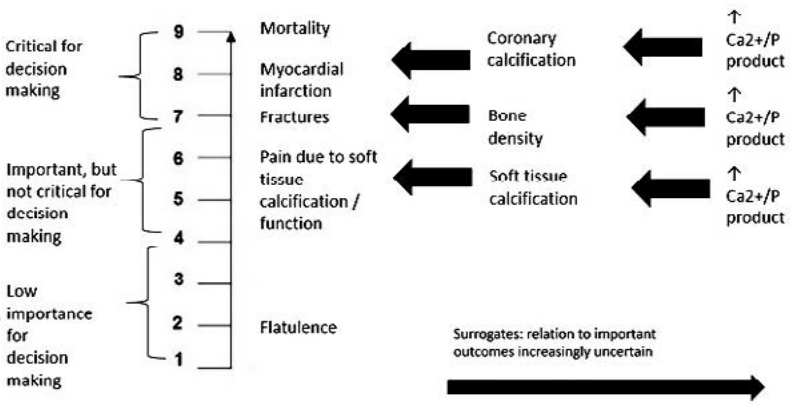

Figure 2. The selection of outcomes during GRADE assessment

of the expert that the impact assessment carried out is acceptable for the approving recommendation. Table 1 shows the classification of quality levels in the GRADE approach and their definition. [2]

Table 2. Significance of levels of evidence

Quality level	Definition
High	We are very confident that the true effect lies close to that of the estimate of the effect
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

During the assessment of the quality of systematic reviews and clinical guidelines the four-point system is used, in which 4 points are for evidence with a high level of quality and 1 point – for evidence with low quality. Initial scores are determined by the study design, RCT are estimated as ones with a high level of accuracy, observational studies, by contrast, have a low level of quality. Then five factors that can lower the total score and three factors that increase scores are taken into account (see Table 3). The GRADE system focuses on the assessment of each important clinical outcome, and the quality of the assessment of different clinical outcomes in the same study or group of studies (meta-analysis) may vary.

Table 3. Criteria for raising and lowering the evidence of the data

Study design	Initial quality of a body of evidence	Lower if	Higher if
Randomized trials	High (4 points)	Risk of Bias -1 Serious -2 Very serious Inconsistency -1 Serious -2 Very serious Indirectness -1 Serious -2 Very serious Imprecision -1 Serious -2 Very serious Publication bias -1 Likely -2 Very likely	Large effect +1 Large +2 Very large Dose response +1 Evidence of a gradient All plausible residual confounding +1 Would reduce a demonstrated effect +1 Would suggest a spurious effect if no effect was observed
	Moderate (3 points)		
Observational studies	Low (2 points)		
	Very low (1 point)		

Risk of bias

The quality of evidence is lowered if there are significant limitations that affect the overall evaluation of therapeutic effect. These limitations include lack of allocation concealment, lack of blinding (especially when the outcomes are subjective), incomplete accounting of patients and outcome events, selective outcome reporting bias, stopping early for benefit. For example, most randomized studies on the relative effectiveness of extensive resection of the tumor in comparison with Whipple's procedure in malignant tumors of the pancreas were limited by lack of optimal level of allocation concealment, lack of prevention of awareness among patients, medical staff and statisticians about which of the observations to which the study groups (experimental or control) and were lost of contact for further observation of a significant number of patients. For these reasons, the accuracy of the data on the most important clinical outcome is estimated as average. [4]

Inconsistency

The inconsistency of effectiveness assessment of the same health technologies in different studies, expressed by the indicators "heterogeneity" or "variability of results", is a signal to reduce the level of quality of evidence on 1 or 2 points depending on the extent of the inconsistency. Fluctuations can arise from differences in populations (differences in severity of disease in different groups), the types of treatment interventions (for example, a more pronounced effect at higher doses), or outcomes (e.g., reduction in treatment effect with time). In this case, the points for inconsistency are lowered. If during the review of meta-analyses are observed: a significant difference in estimate of effect, the overlap of confidence intervals, statistical tests of heterogeneity of results with $p < 0$ (test of the null hypothesis), the high value of the I^2 (Index of heterogeneity) (>75%), the points for inconsistency are lowered. [7]

Indirectness

The main traits of indirectness in systematic review are the differences in compared populations, types of treatment, clinical outcomes. The GRADE approach prefers a direct comparison of health technology in comparison with indirect comparisons [8]

Imprecision

The error in the determination of effect is found when in the studies with a relatively small number of patients and small number of events wide confidence interval of the estimate of effect is observed. One more indicator of imprecision is optimal information size. If the total number of patients included in a systematic review is less than the number of patients modelled in the calculations of the size of the conditional sampling for a statistically correct study, it is necessary to reduce the points for imprecision.[6]

Table 4. Optimal information size determination

Total number of events	Relative Risk Reduction	Implications for meeting Optimal information size implications threshold
100 or less	<30%	Will almost never meet threshold whatever control event rate
200	30%	Will meet threshold for control event rates for 25% or greater
200	25%	Will meet threshold for control event rates for 50% or greater
200	20%	Will meet threshold for control event rates for 80% or greater
300	≥ 30%	Will meet threshold
300	25%	Will meet threshold for control event rates for 25% or greater
300	20%	Will meet threshold for control event rates for 60% or greater
400 or more	≥ 25%	Will meet threshold for any control event rate
400 or more	20%	Will meet threshold for control event rates for 40% or greater

Publication bias

Publication bias arises because of the tendency of some researchers, editors, and other persons to publish mostly positive (statistically significant) results of scientific research, omitting statistically insignificant, ambiguous or contradictory expectations data. It is always necessary to suspect this kind of error in the small "positive" studies, sponsored research, the systematic publication of "positive" or "negative" effects with a small sample. To identify publication errors GRADE uses the funnel plots. In the absence of a publication error the graph has the form of a symmetrical funnel (Fig.3) If there is a high risk of publishing papers that have statistically significant results in favor of the intervention being evaluated, in the lower left corner of the funnel there is a gap in place of the missing results, hence there is a need to reduce the score. [5]

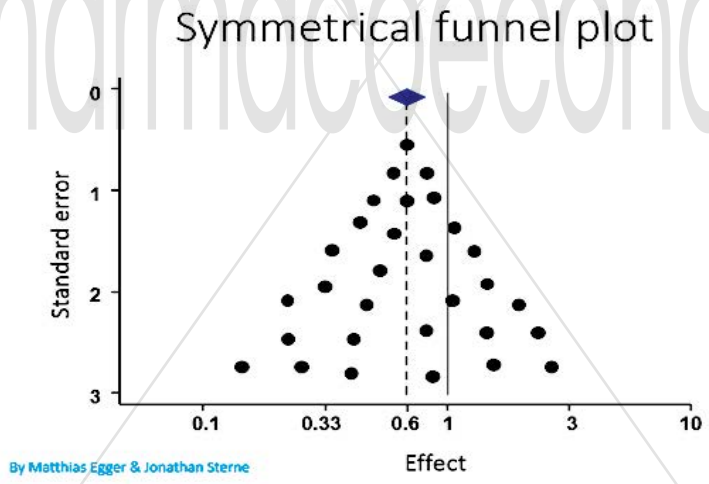

Figure 3. Symmetrical funnel plot

Factors rating up the quality of evidence

Despite the fact that typically observational studies are evaluated as evidence of low level of quality, under certain conditions their level of quality can be improve. If methodologically right built the study provides a significant size effect or if there is evidence of a dose-dependent effect, we can be confident in the quality of the study and depending on the severity of the factors is recommended to improve assessment of quality of 1 or 2 points. An example of increasing the reliability of the data can serve as a meta-analysis that included observational studies of the effect of prophylaxis with warfarin in heart valve replaced. This study showed that the relative risk of thromboembolism prophylaxis with warfarin was 0.17 (95% confidence interval 0.13-0.24). Very high values of effect size indicate a high level of data reliability, therefore, should increase the score by 2 points. An example of increasing the reliability of the data can serve is a meta-analysis that included observational studies of the effect of prophylaxis with warfarin in patients with heart valve replaced. This study showed that the relative risk of thromboembolism prophylaxis with warfarin was 0.17 (95% confidence interval 0.13-0.24). Very high values of effect size indicate a high level of data reliability, therefore, should increase the score by 2 points. [9]

Strength of recommendations

The strength of the recommendations reflects the degree of confidence of the expert that positive effects from this therapy exceed its shortcomings. Positive effects include reduction of mortality, increase in the quality of life, reducing the burden of disease and reduced use of health resources. Undesirable effects in this context include side effects that affect the deterioration of the positive effects. Binary system, evaluating the strength of recommendations as "strong" and "weak" is used.

Assessment of the strength of recommendations can be performed from the perspective of patients, physicians and decision makers. For each group, the definition of "strong" recommendation is as follows:

- For patients—most people in your situation would want the recommended course of action and only a small proportion would not; request discussion if the intervention is not offered
- For clinicians—most patients should receive the recommended course of action
- For policy makers—the recommendation can be adopted as a policy in most situations.

The implications of a weak recommendation are:

- For patients—most people in your situation would want the recommended course of action, but many would not
- For clinicians—you should recognize that different choices will be appropriate for different patients and that you must help each patient to arrive at a management decision consistent with her or his values and preferences
- For policy makers—policy making will require substantial debate and involvement of many stakeholders.

Factors, determining the strength of recommendations

Making decisions on the degree of strength of recommendations, the expert group considers four key factors:

- quality of evidence;
- balance between desirable and undesirable effects;
- values and preferences;
- costs

The first key factor in determining the strength of recommendations is the balance between the positive effects and unwanted effects of the use of different methods of treatment. For example, consider the use of antenatal steroids in preterm delivery. The acceptance of steroids by the mother reduces the risk of respiratory distress syndrome in the newborn with minimal increased risk of adverse events, increasing costs and is accompanied by discomfort in the use of drugs. The benefits of therapy significantly outweigh the disadvantages, therefore, the strength of the recommendation of this scheme of treatment is determined as strong. (a high degree of credibility of the recommendation of this regimen). When the advantages and disadvantages of therapy are approximately equivalent, then such recommendation should be regarded as weak (with low credibility). For example, consider a population of patients with atrial fibrillation and a low stroke risk. Warfarin can more effectively reduce the risk of stroke, but its use is accompanied by increased risk of bleeding and inconvenience in use. In this case, a better strategy is to consider each case individually. The second factor is the reliability of scientific data. If the experts are not sure about the effect size and adverse effects associated with taking a new drug, the recommendation cannot be strong. For instance, graduated compression stockings have an apparent large effect in reducing deep venous thrombosis in people making long plane journeys. The randomized trials from which the estimate of effect comes were, however, seriously flawed—the techniques for measuring deep venous thrombosis were not reproducible, and the studies were unblinded. Despite the apparent large benefit, use of stockings warrants only a weak recommendation. Taking into account that compared health technologies are the characteristic advantages and disadvantages, how the expert group identifies the advantages, convenience, and risk from the use of new technologies and whether these assessments with the opinion of patients is vital for making decisions about the credibility of the recommendations.

Consider the subject of preventing strokes in patients with atrial fibrillation. Warfarin, relative to no antithrombotic therapy, reduces the risk of stroke by approximately 65% but increases the risk of severe gastrointestinal bleeding. Devereaux and colleagues asked 63 physicians and 61 patients how many serious gastrointestinal bleeds they would tolerate in 100 patients and still be willing to prescribe or take warfarin to prevent eight strokes (four minor and four major) in 100 patients. Whereas physicians gave a wide diversity of responses, most patients placed a high value on avoiding a stroke and were ready to accept a bleeding risk of 22% to reduce their chances of having a stroke by 8%. Even among patients, however, diversity in values and preferences was apparent; a few patients were ready to accept only a small risk of bleeding. These data suggest that only in patients at high risk of stroke would a strong recommendation for warfarin be warranted [11]. A very important factor in determining the strength of the recommendations is the economic effect of the application of the estimated medical technology. For this factor, the most characteristic variability depending on time and region. When you migrate data from one country to another must take account of differences in the organization and funding of health systems, differences in the cost structure in real clinical practice. It should be noted that a conditional recommendation should not be an obstacle for further research in the field of advanced medical technology. In particular, if the expert board evaluates innovative health technologies at an acceptable cost, a satisfactory correlation between the positive effects and adverse events, with encouraging clinical results, but the reliability of evidence is low, we should accept the recommendation conditional, limiting the technology only for research purposes. This policy adheres to NICE, giving the opportunity to stakeholders to continue working to improve the evidence base.

As an example of the application of the GRADE approach, we will give a part of report of the Institute of public health of Norway (Kunnskapssenteret) on assessment of health technologies used in the treatment of multiple sclerosis. [12] Two researchers conducted an independent assessment of used meta-analyses, to determine their accuracy for each of the selected clinical outcome. Clinical outcomes were selected by a group of doctors using the PICO: annualised relapse rate, the level of progression, and withdrawal due to adverse events. Next, for each of the compared treatment regimens were evaluated the quality of the data on the selected outcome, taking into account risk of bias, inconsistency of results, the indirectness of data, imprecision of the definition of effect size, risk of bias related to the preferred publication of positive results of a study size effect, the dependence of the effect of dose and unaccounted factors, the exclusion of which reduces the size of the found effect. As a result, for the drug Interferon beta-1a 30 mcg was built table for

Table 4. Table of evidence for Interferon beta-1a 30 mcg compared with placebo

Quality assessment							№ of patients		Effect			
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Interferon beta-1a 30 mcg	Placebo	Relative (95% CI)	Absolute (95% CI)	Quality	Importance
Annualised relapse rate												
3	randomised trials	not serious	not serious	not serious [1]	not serious	none	-/659	-/647	**RR 0.76** (0.65 to 0.89)	0 fewer per 1000 (from 0 fewer to 0 fewer)	⊕⊕⊕⊕ HIGH [1]	
Disease Progression												
2	randomised trials	not serious	not serious	not serious [1]	serious [2]	none	70/605 (11.6%)	96/593 (16.2%)	**RR 0.68** (0.50 to 0.95)	52 fewer per 1000 (from 8 fewer to 81 fewer)	⊕⊕⊕○ MODERATE [12]	
Withdrawal due to adverse events												
3	randomised trials	not serious	not serious	not serious [1]	very serious [2,3]	none	34/659 (5.2%)	21/647 (3.2%)	**RR 1.73** (0.82 to 3.87)	24 more per 1000 (from 6 fewer to 93 more)	⊕⊕○○ LOW [1,2,3]	

MD – mean difference, RR – relative risk

table on the assessment of evidence (Evidence profile) (table 4).

According to the results of the evaluation, we can say that evidence on the annualised relapse rate in a year while taking Interferon beta-1a 30 mcg compared to placebo have a high degree of quality. Data on the level of disease progression while receiving Interferon beta-1a 30 mcg compared with a placebo have an average level of quality. Information about the frequency of withdrawal due to adverse events while taking Interferon beta-1a 30 mcg compared to placebo have a low degree of quality.

Conclusion

The GRADE system offers a transparent and comprehensible method of assessing evidence, given the methods of modern evidence-based medicine. The application of this method is a step towards the adoption of more rational, informed decisions, taking into account the opinions of patients, doctors and decision-makers.

Refererences

1. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383394. doi: 10.1016/j.jclinepi.2010.04.026
2. Guyatt GH, Oxman AD, Kunz R, et al. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008;336(7651):995-998. doi: 10.1136/bmj.39490.551019.BE
3. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011;64(4):395-400. doi: 10.1016/j.jclinepi.2010.09.012
4. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011;64(4):407-415. doi: 10.1016/j.jclinepi.2010.07.017
5. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011;64(12):1277-1282. doi: 10.1016/j.jclinepi.2011.01.011
6. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011;64(12):1283-1293. doi: 10.1016/j.jclinepi.2011.01.012
7. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011;64(12):1294-1302. doi: 10.1016/j.jclinepi.2011.03.017
8. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011;64(12):1303-1310. doi: 10.1016/j.jclinepi.2011.04.014
9. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-1316. doi: 10.1016/j.jclinepi.2011.06.004
10. Guyatt G, Oxman AD, Sultan S, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol. 2013;66(2):151-157. doi: 10.1016/j.jclinepi.2012.01.006
11. Guyatt Gordon H, Oxman Andrew D, Kunz Regina, Falck-Ytter Yngve, Vist Gunn E, Liberati Alessandro et al. Going from evidence to recommendations BMJ 2008; 336 :1049
12. Amatya B, Khan F, La Mantia L, Demetrios M, Wade DT. Non pharmacological interventions for spasticity in multiple sclerosis. Cochrane Database Syst Rev. 2013;2:CD009974. doi: 10.1002/14651858.CD009974.pub2