Supplementary Table 1: Qualitative overview of different grading systems (please see subsequent tables for more details). GRADE: Grading of Recommendations Assessment, Development and Evaluation ([Guyatt, Oxman, Kunz, et al. 2008](#_ENREF_6); [Berkman et al. 2015](#_ENREF_3); [GRADE 2017](#_ENREF_5)); CEBM: Oxford Centre for Evidence-based Medicine ([CEBM 2009](#_ENREF_4)); SIGN: Scottish Intercollegiate Guidelines Network ([SIGN 2013](#_ENREF_14), [2014](#_ENREF_15)); AHRQ: Agency for Healthcare Research and Quality ([Berkman et al. 2015](#_ENREF_3)), AHCPR: Agency for Health Care Policy and Research ([AHCPR 1992](#_ENREF_2); [Hadorn et al. 1996](#_ENREF_8)), USPSTF: U.S. Preventive Services Task Force ([USPSTF 2012](#_ENREF_16)); NHMRC: National Health and Medical Research Council ([NHMRC 2009](#_ENREF_9)).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GRADE** | **SIGN** | **AHRQ** | **USPSTF** | **OCEBM** | **AHCPR** | **NHMRC** |
| **Levels of evidence** | High Moderate  Low  Very low | 1++  1+  1-  2++  2+  2-  3  4 | High  Moderate  Low  Insufficient | High Moderate  Low | 1a  1b  1c  2a  2b  2c  3a  3b  4  5 | Ia  Ib  IIa  IIb  III  IV | I  II  III-1  III-2  III-3  IV |
| **Grades of recommendations/**  **Strength of recommendations** | Strong yes  Weak yes  Strong no  Weak no | A  B  C  D  Good practice points |  | A  B C D I | A  B  C  D | A  B C  (D) | A  B  C  D |

Supplementary Table 2: GRADE apporach for grading evidence ([Guyatt, Oxman, Vist, et al. 2008](#_ENREF_7); [Berkman et al. 2015](#_ENREF_3); [GRADE 2017](#_ENREF_5))

|  |  |  |
| --- | --- | --- |
| **Quality of evidence** | | |
| High quality | Further research is very unlikely to change our confidence in the estimate of effect | Randomized trials without serious limitations Well-performed observational studies with very large effects (or other qualifying factors) |
| Moderate quality | Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate | Randomized trials with serious limitations Well-performed observational studies yielding large effects |
| Low quality | Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate | Randomized trials with very serious limitations Observational studies  without special strengths or important limitations |
| Very low quality | Any estimate of effect is very uncertain | Randomized trials with very serious limitations and inconsistent results. Observational studies with serious limitations. Unsystematic clinical observations (e.g. case series or case reports) |
|  | |  |
| **Strength of recommendations** | | |
|  | Strong recommendation for using an intervention | |
|  | Weak recommendation for using an intervention | |
|  | Weak recommendation against using an intervention | |
|  | Strong recommendation against using an intervention | |
|  | | |
| **Determinants of strength of recommendation** | | |
| Balance between desirable and undesirable effects | The larger the difference between the desirable and undesirable effects, the higher the likelihood that a strong recommendation is warranted. The narrower the gradient, the higher the likelihood that a weak recommendation is warranted | |
| Quality of evidence | The higher the quality of evidence, the higher the likelihood that a strong recommendation is warranted | |
| Values and preferences | The more values and preferences vary, or the greater the uncertainty in values and preferences, the higher the likelihood that a weak recommendation is warranted | |
| Costs (resource allocation) | The higher the costs of an intervention—that is, the greater the resources consumed—the lower the likelihood that a strong recommendation is warranted | |

Supplementary Table 3: SIGN grading system: The grade of recommendation relates to the strength of the evidence on which the recommendation is based. It does not reflect the clinical importance of the recommendation ([SIGN 2013](#_ENREF_14)).

|  |  |
| --- | --- |
| **Levels of evidence** | |
| 1++ | High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias |
| 1+ | Well conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias |
| 1- | Meta-analyses, systematic reviews, or RCTs with a high risk of bias |
| 2++ | High quality systematic reviews of case control or cohort studies; High quality case control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal |
| 2+ | Well conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal |
| 2 | Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal |
| 3 | Non-analytic studies, e.g. case reports, case series |
| 4 | Expert opinion |
|  | |
| **Grades of Recommendations** | |
| A | At least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population;  or  A body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating overall consistency of results |
| B | A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results;  or  Extrapolated evidence from studies rated as 1++ or 1+ |
| C | A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results; or  Extrapolated evidence from studies rated as 2++ |
| D | Evidence level 3 or 4;  or  Extrapolated evidence from studies rated as 2+ |
|  | |
| **Good practice points** | Recommended best practice based on the clinical experience of the guideline development group |

Supplementary Table 4: US Agency for Healthcare Research and Quality (AHRQ) strength of evidence grades and definitions ([Owens et al. 2009](#_ENREF_12); [Berkman et al. 2015](#_ENREF_3))

|  |  |
| --- | --- |
| **Grade** | **Definition** |
| **High** | **We are very confident that the estimate of effect lies close to the true effect for this outcome**. The body of evidence has few or no  deficiencies. We believe that the findings are stable, that is, another study would not change the conclusions. |
| **Moderate** | **We are moderately confident that the estimate of effect lies close to the true effect for this outcome.** The body of evidence has some  deficiencies. We believe that the findings are likely to be stable, but some doubt remains. |
| **Low** | **We have limited confidence that the estimate of effect lies close to the true effect for this outcome.** The body of evidence has major or  numerous deficiencies (or both). We believe that additional evidence is needed before concluding either that the findings are stable  or that the estimate of effect is close to the true effect. |
| **Insufficient** | **We have no evidence, we are unable to estimate an effect, or we have no confidence in the estimate of effect for this outcome**. No evidence is available or the body of evidence has unacceptable deficiencies, precluding reaching a conclusion |

**Supplementary Table 5:** U.S. Preventive Services Task Force (USPSTF) grading levels ([USPSTF 2012](#_ENREF_16))

|  |  |
| --- | --- |
| **Levels of certainty** | |
| High | The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies. |
| Moderate | The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as:   * The number, size, or quality of individual studies. * Inconsistency of findings across individual studies. * Limited generalizability of findings to routine primary care practice. * Lack of coherence in the chain of evidence.   As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusions |
| Low | The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of:   * The limited number or size of studies. * Important flaws in study design or methods. * Inconsistency of findings across individual studies. * Gaps in the chain of evidence. * Findings not generalizable to routine primary care practice. * Lack of information on important health outcomes. More information may allow estimation of effects on health outcomes. |
|  | |
| **Grades of Recommendations** | |
| A | The USPSTF recommends the service. There is high certainty that the net benefit is substantial |
| B | The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. |
| C | Clinicians may provide this service to selected patients depending on individual circumstances. However, for most individuals without signs or symptoms there is likely to be only a small benefit from this service (statement under revision) |
| D | The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. |
| I | The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined. |

Supplementary Table 6: Oxford Centre for Evidence-based Medicine (OCEBM) levels of evidence ([CEBM 2009](#_ENREF_4))

|  |  |
| --- | --- |
| **Levels of evidence** | |
| 1a | Systematic review/Meta-Analysis (with homogeneity) of RCTs |
| 1b | Individual RCT (with narrow confidence Interval) |
| 1c | All or none |
| 2a | Systematic review/Meta-Analysis (with homogeneity) of cohort studies |
| 2b | Individual cohort study (including low quality RCT; e.g., <80% follow-up) |
| 2c | “Outcomes” Research; Ecological studies |
| 3a | Systematic review/Meta-Analysis (with homogeneity) of case-control studies |
| 3b | Individual Case-Control Study |
| 4 | Case-series (and poor quality cohort and case-control studies) |
| 5 | Expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles |
|  | |
| **Grades of Recommendations** | |
| A | Consistent level 1 studies |
| B | Consistent level 2 or 3 studies or extrapolations from level 1 studies |
| C | Level 4 studies or extrapolations from level 2 or 3 studies |
| D | Level 5 evidence or troublingly inconsistent or inconclusive studies of any level |

Supplementary Table 7: Agency for Health Care Policy and Research grading system (AHCPR) ([AHCPR 1992](#_ENREF_2); [Hadorn et al. 1996](#_ENREF_8)). This system has been introduced in the early 1990 and has been used for years in guidelines across all medical fields. Many features laid down in this system can be found in newer grading systems.

|  |  |
| --- | --- |
| **Levels of evidence** | |
| Ia | Meta-analysis of randomized controlled trials |
| Ib | At least one randomized controlled trial |
| IIa | At least one well-designed controlled study without randomization |
| IIb | At least one other type of well-designed quasi-experimental study |
| III | Well-designed non-experimental descriptive studies, such as comparative studies, correlation studies or case-control studies |
| IV | Expert opinions and/or clinical experience of respected authorities |
|  | |
| **Grades of Recommendations** | |
| A | Directly based on category I evidence |
| B | Directly based on category II evidence or extrapolated from category I evidence |
| C | Directly based on category III evidence or extrapolated from category I or II evidence |
| (D) | (Directly based on category IV evidence or extrapolated from category I, II or III evidence) |
| *Alternative* |  |
| A | Requires at least one randomized controlled trial as part of a body of literature of overall good quality and consistency addressing the specific recommendations (Ia, Ib) |
| B | Require availability of well-conducted clinical studies bit no randomized clinical trials on the topic of recommendations (IIa, IIb, III) |
| C | Require evidence obtained from expert committee reports or opinions and/or clinical experiences of respected authorities. Indicates an absence of directly applicable clinical studies of good quality (IV) |

Supplementary Table 8: Australian National Health and Medical Research Council (NHMRC) levels of evidence ([NHMRC NHaMRC 2009](#_ENREF_10))

|  |  |
| --- | --- |
| **Levels of evidence** | |
| I | A systematic review of level II studies |
| II | A randomised controlled trial |
| III-1 | A pseudo-randomised controlled trial (i.e. alternate allocation or some other method) |
| III-2 | Acomparative study with concurrent controls:   * Non-randomised, experimental trial * Cohort study * Case-control study * Interrupted time series with a control group |
| III-3 | A comparative study without concurrent controls:   * Historical control study * Two or more single-arm studies * Interrupted time series without a parallel control group |
| IV | Case series with either post-test or pre-test/posttest outcomes |
| **Grades of recommendation** | |
| A | Body of evidence can be trusted to guide practice |
| B | Body of evidence can be trusted to guide practice in most situations |
| C | Body of evidence provides some support for recommendation(s) but care should be taken in its application |
| D | Body of evidence is weak and recommendation must be applied with caution |

Supplementary Table 9: Strength of Recommendation System for NICE Guidance ([Addington et al. 2017](#_ENREF_1); [NICE 2017](#_ENREF_11); [Pringsheim and Addington 2017](#_ENREF_13)).

|  |
| --- |
| **Strength of recommendations:** The wording used in NICE recommendations denotes the certainty with which the recommendation is made (the strength of the recommendation) |
| **Interventions that must (or must not) be used:** We usually use ‘must’ or ‘must not’ only if there is a legal duty to apply the recommendation. Occasionally we use ‘must’ (or ‘must not’) if the consequences of not following the recommendation could be extremely serious or potentially life threatening |
| **Interventions that should (or should not) be used—a ‘strong’ recommendation:** We use ‘offer’ (and similar words such as ‘refer’ or ‘advise’) when we are confident that, for the vast majority of patients, an intervention will do more good than harm and will be cost effective. |
| **Interventions that could be used:** We use ‘consider’ when we are confident that an intervention will do more good than harm for most patients, and will be cost effective, but other options may be similarly cost effective. The choice of intervention, and whether to have the intervention at all, is more likely to depend on the patient’s values and preferences than for a strong recommendation. |

**References**

Addington D, Abidi S, Garcia-Ortega I, Honer WG, Ismail Z. 2017. Canadian Guidelines for the Assessment and Diagnosis of Patients with Schizophrenia Spectrum and Other Psychotic Disorders. Canadian journal of psychiatry Revue canadienne de psychiatrie. 62(9):594-603.

AHCPR. 1992. Acute pain management: operative or medical procedures and trauma. Rockville MD: Agency for Health Care Policy and Research Publications.

Berkman ND, Lohr KN, Ansari MT, Balk EM, Kane R, McDonagh M, Morton SC, Viswanathan M, Bass EB, Butler M et al. 2015. Grading the strength of a body of evidence when assessing health care interventions: an EPC update. Journal of clinical epidemiology. 68(11):1312-1324.

CEBM. 2009. Oxford Centre for Evidence-based Medicine. [accessed 30.12.2017]. <http://www.cebm.net/blog/2009/06/11/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/>.

GRADE. 2017. Grading of Recommendations Assessment, Development and Evaluation. [accessed 30.12.2017]. <http://gradeworkinggroup.org/>.

Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schunemann HJ, Group GW. 2008. Going from evidence to recommendations. Bmj. 336(7652):1049-1051.

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ, Group GW. 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Bmj. 336(7650):924-926.

Hadorn DC, Baker D, Hodges JS, Hicks N. 1996. Rating the quality of evidence for clinical practice guidelines. Journal of clinical epidemiology. 49(7):749-754.

NHMRC. 2009. NHMRC additional levels of evidence and grades for recommendations for developers of guidelines. [accessed 30.12.2017]. https://[www.nhmrc.gov.au/\_files\_nhmrc/file/guidelines/developers/nhmrc\_levels\_grades\_evidence\_120423.pdf](http://www.nhmrc.gov.au/_files_nhmrc/file/guidelines/developers/nhmrc_levels_grades_evidence_120423.pdf).

NHMRC NHaMRC. 2009. NHMRC Additional Levels of Evidence and Grades for Recommendations for Developers of Guidelines. Melbourne.

NICE TNIfHaCE. 2017. Developing NICE guidelines: the manual. [accessed 27.12.2017]. https://[www.nice.org.uk/process/pmg20/chapter/developing-and-wording-recommendations-and-writing-the-guideline](http://www.nice.org.uk/process/pmg20/chapter/developing-and-wording-recommendations-and-writing-the-guideline).

Owens DK, Lohr KN, Atkins D. 2009. Grading the strength of a body of evidence when comparing medical interventions. In: Agency for Healthcare Research and Quality. Methods Guide for Comparative Effectiveness Reviews

Pringsheim T, Addington D. 2017. Canadian Schizophrenia Guidelines: Introduction and Guideline Development Process. Canadian journal of psychiatry Revue canadienne de psychiatrie. 62(9):586-593.

SIGN SIGN. 2013. SIGN 131 • Management of schizophrenia • A national clinical guideline.

SIGN SIGN. 2014. Critical appraisal notes and checklists. [accessed 27.12.2017]. <http://www.sign.ac.uk/checklists-and-notes.html>.

USPSTF. 2012. U.S. Preventive Services Task Force (USPSTF) grading system. [accessed 30.12.2017]. https://[www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions](http://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions).