Clinical Chemistry 54:11
1872–1882 (2008)

Do Guidelines for the Diagnosis and Monitoring of Diabetes Mellitus Fulfill the Criteria of Evidence-Based Guideline Development?

Eva Nagy,[1] Joseph Watine,[2] Peter S. Bunting,[3] Rita Onody,[1] Wytze P. Oosterhuis,[4] Dunja Rogic,[5] Sverre Sandberg,[6] Krisztina Boda,[7] and Andrea R. Horvath[1*]

BACKGROUND: Although the methodological quality of therapeutic guidelines (GLs) has been criticized, little is known regarding the quality of GLs that make diagnostic recommendations. Therefore, we assessed the methodological quality of GLs providing diagnostic recommendations for managing diabetes mellitus (DM) and explored several reasons for differences in quality across these GLs.

METHODS: After systematic searches of published and electronic resources dated between 1999 and 2007, 26 DM GLs, published in English, were selected and scored for methodological quality using the AGREE Instrument. Subgroup analyses were performed based on the source, scope, length, origin, and date and type of publication of GLs. Using a checklist, we collected laboratory-specific items within GLs thought to be important for interpretation of test results.

RESULTS: The 26 diagnostic GLs had significant shortcomings in methodological quality according to the AGREE criteria. GLs from agencies that had clear procedures for GL development, were longer than 50 pages, or were published in electronic databases were of higher quality. Diagnostic GLs contained more preanalytical or analytical information than combined (i.e., diagnostic and therapeutic) recommendations, but the overall quality was not significantly different. The quality of GLs did not show much improvement over the time period investigated.

CONCLUSIONS: The methodological shortcomings of diagnostic GLs in DM raise questions regarding the validity of recommendations in these documents that may affect their implementation in practice. Our results suggest the need for standardization of GL terminology and for higher-quality, systematically developed recommendations based on explicit guideline development and reporting standards in laboratory medicine.

Clinical practice guidelines are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances *(1)*. The methodological quality of practice guidelines (GLs)[8] has been widely criticized *(2)*. As effective treatment requires effective diagnosis, recommendations for the clinical use of tests should also fulfill the criteria of evidence-based guideline development *(3)*. Assuring methodological quality of GLs requires that the potential biases of GL development have been addressed adequately and that the recommendations are valid and feasible in practice. Therefore, the aim of our current study was to investigate whether GL development teams use appropriate and explicit methods for making diagnostic recommendations and whether diagnostic GLs meet basic reporting standards. For these assessments, we chose laboratory diagnosis and monitoring of diabetes mellitus (DM), one of the global health problem areas in which GLs are most widely used worldwide. In previous work using the AGREE (Appraisal of Guidelines Research and Evaluation) instrument *(4)*, we showed that the methodological quality of 4 DM GLs, issued by presti-

[1] Department of Clinical Chemistry, University of Szeged, Medical Faculty, Szeged, Hungary; [2] Laboratoire de Biologie Polyvalente, Hôpital Général, Rodez, France; [3] Department of Pathology and Laboratory Medicine, The Ottawa Hospital, Ottawa, Ontario, Canada; [4] Department of Clinical Chemistry, Atrium Medical Centre, Heerlen, The Netherlands; [5] Institute of Clinical Laboratory Diagnosis, Zagreb University School of Medicine, Clinical Hospital Center, Zagreb, Croatia; [6] Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen, Norway; [7] Department of Medical Informatics, University of Szeged, Medical Faculty, Szeged, Hungary.

* Address correspondence to this author at: Department of Clinical Chemistry, University of Szeged, Medical Faculty, Somogyi Bela ter 1, Szeged, H-6725 Hungary. E-mail ahorvath@clab.szote.u-szeged.hu.

Received August 6, 2008; accepted August 7, 2008.
Previously published online at DOI: 10.1373/clinchem.2008.109082

[8] Nonstandard abbreviations: GL, guideline; DM, diabetes mellitus; AGREE, Appraisal of Guidelines Research and Evaluation; D, AGREE domain; I, AGREE item.

gious authorities between 1999 and 2003, was rather low *(5)*. The objectives of the current study were to examine whether similar findings were present in a larger sample of recently published DM GLs, and if so, to identify some of the reasons for such findings. We were also interested in differences between GLs that are primarily diagnostic compared to those that are combined with therapeutic recommendations. Owing to the high number and heterogeneous nature of diagnostic questions and recommendations addressed in DM guidelines, and the fact that AGREE is designed to assess the methodological quality of GLs only, our study did not investigate the accuracy of the content of guidelines.

Materials and Methods

SEARCH STRATEGY

We carried out a systematic literature search to retrieve diagnostic GLs in DM. The aim of the search was to obtain a representative sample of GLs, published in English between 1 January 1999 and 31 December 2007, that can be easily accessed and are therefore likely to be read and used in many countries. In PubMed, 1 reviewer (J. Watine) applied a broad search strategy using the Clinical Queries filter "systematic[sb]," which is capable of retrieving systematic reviews and/or GLs *(6)*. This term was combined with the laboratory-specific MeSH terms "Clinical Laboratory Techniques" [MeSH] AND systematic[sb] OR "Laboratory Techniques and Procedures"[MeSH] AND systematic[sb] *(7)*. Another independent reviewer (E. Nagy) searched in electronic journals using the keywords "guideline" AND "diabetes," and in dedicated GL databases and websites of professional organizations. The databases searched are shown in Data Supplement 1, which accompanies the online version of this article at http://www.clinchem.org/content/vol54/issue11.

SELECTION OF GUIDELINES ELIGIBLE FOR THE STUDY

Based on the titles and/or abstracts, references were screened for relevance (E. Nagy, J. Watine). Using this subset, 2 independent reviewers (E. Nagy, A.R. Horvath) applied the following inclusion criteria: the publication fulfilled the definition of GLs *(1)* and dealt with the use of laboratory tests for the diagnosis or monitoring of DM, and the GL was publicly available in a peer-reviewed journal and/or in nationally or internationally endorsed GL databases. If several updates of the GL were available during the studied time period, only the latest version was selected. All these criteria had to be met for enrollment into the study.

We excluded publications that contained therapeutic recommendations only, were primarily focused on technical/analytical/quality control/standardization/quality management issues, referred to special patient groups, or offered local protocols on best practice (i.e., restricted to 1 particular health care setting).

EVALUATION OF THE METHODOLOGICAL QUALITY OF GUIDELINES

It has been shown that compliance with the AGREE criteria of most GL development programs is high *(8)*, and, therefore, we used AGREE, a standardized, generic, and validated checklist *(4, 9, 10)*, along with its accompanying Training Manual, for the assessment of GLs. AGREE arranges 23 criteria, thereafter referred to as AGREE items I1 through I23, into 6 key domains (D1 through D6). Selected GLs were randomly allocated to 2 assessment teams with 4 reviewers per team. Reviewers were trained how to use AGREE in a pilot study before conducting this larger survey *(5)*. To make the appraisal process as objective as possible, reviewers were provided with all supplementary files referenced by each GL and found in the public domain (E. Nagy), including background supporting materials, technical papers, or general GL development manuals issued by the respective GL agency.

Reviewers independently assessed the fulfillment of the AGREE criteria on a 4-point Likert scale. Disagreements in 2 or more scores between appraisers were resolved by discussion and consensus. An independent reviewer's opinion (A.R. Horvath) was required in 1 case only for reaching consensus. Domain scores were expressed in percentages, and a final conclusion was reached about the acceptability of the GL according to the instructions of AGREE. A GL was "strongly recommended" if the majority of items scored 3 or 4 and most domain scores (i.e., at least 4 of 6) were >60%. A GL was "not recommended" if the majority of items rated 1 or 2 and most of the domain scores (i.e., 4 or more of 6) were <30%. Guidelines were "recommended with provisos or alterations" when the GL rated high (3 or 4) or low (1 or 2) on a similar number of items and most domain scores were between 30% and 60%. For investigating whether diagnostic GLs meet additional reporting standards *(11)* that are not covered in depth in the AGREE, we assessed the presence of the following items: (*a*) an evidence table, (*b*) a description of the grading system, (*c*) graded recommendations, and (*d*) an expiry or review date. Additionally, we assessed whether the GL contained data thought to be important for test interpretation *(3)*, such as (*e*) prevalence, (*f*) diagnostic accuracy of tests, (*g*) preanalytical, and (*h*) analytical specifications. All reviewers checked the availability of these items, and results were summarized by 1 independent assessor (E. Nagy).

We created 5 subgroups of GLs based on their source, scope, length, and origin and whether they were

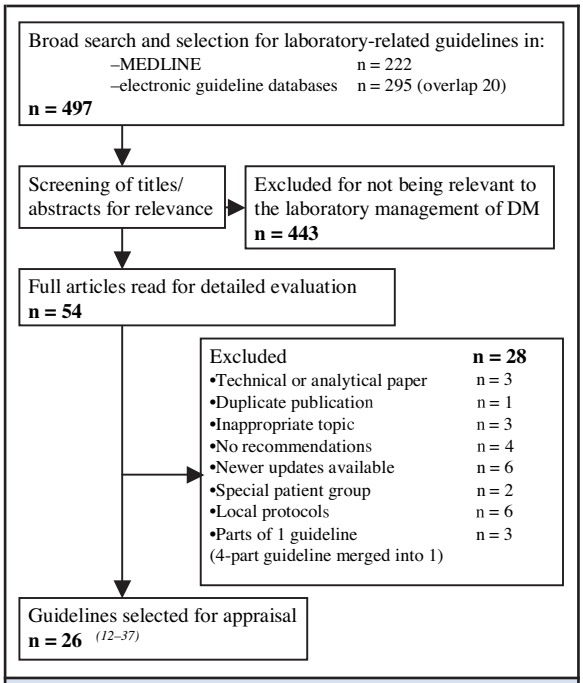

Fig. 1. Searching and selecting guidelines.

supplemented with a guideline methods manual. We also investigated the quality of guidelines according to the date and type of publication. In the statistical analyses (K. Boda), the mean item and standardized domain scores of GL subgroups were compared by the Kruskal–Wallis test. Pair-wise comparisons were carried out using the Mann–Whitney *U*-test with Bonferroni correction. The frequency of reporting laboratory specific information in different guideline subgroups was compared with the Fisher exact test. The level of significance was set at $P \leq 0.01$ because of multiple comparisons. All analyses were performed using SPSS for Windows, version 13.

Results

Of 2630 references retrieved in a broad search, we found 497 GLs to be related to laboratory medicine (Fig. 1). After screening for relevance, we subjected 54 GLs to the selection criteria described. Fig. 1 shows the reasons for excluding 28 GLs; 26 GLs became eligible for critical appraisal *(12–37)*. The most important characteristics of GLs are summarized in Table 1. Nine GLs originated from the USA, 3 from Canada, 7 from the UK, 1 each from Australia, New Zealand, and South Africa, and 4 were international. All but 2 GLs were developed in the last 6 years.

CRITICAL APPRAISAL OF GUIDELINES

Based on the assessment of methodological quality, 22 GLs were recommended by reviewers, of which only 11 were strongly recommended and the rest "with provisos and alterations." Four GLs had 4 or 5 domains with scores <30%, and reviewers did not recommend their use (Table 1).

The domain and item scores of individual GLs are shown in Table 1 and online Data Supplement 2, respectively. Table 2 summarizes the mean item scores and the number and proportion of GLs scoring >3 on the 4-point Likert scale. Overall, the best-performing domains were D1, "scope and purpose" (77%; Table 1), with a high proportion of GLs scoring above 3 for all items (Table 2). Although D4, "clarity and presentation," scored highly (76%; Table 1), I18 within this domain performed poorly, as only 10 GLs (38%) were supported with tools for application (Table 2).

Domains 2 and 3, which explored the process of GL development, showed lower scores (Table 1). Nine GLs (35%) scored >60% in "stakeholder involvement" and 14 (54%) in "rigor of development" domains. Of individual items in D2, only a small proportion of GLs gave information about the composition and affiliations of the guideline development group, provided some information on patient involvement in the development process, defined their target users clearly, and pilot tested the GL by target users before publication (Table 2). In D3, there are notable shortcomings in using systematic methods for searching the evidence or at least giving some information about literature retrieval, describing clearly the criteria for selecting the evidence, indicating the methods used for formulating recommendations, and giving information on the peer review and updating process. The lowest scores were achieved with "applicability" (34%) and "editorial independence" (39%) domains, in which each items performed very poorly (Table 2 and online Data Supplement 2).

QUALITATIVE ANALYSIS OF GUIDELINES

Date of publication. Quality of GLs was also investigated according to the date of publication to see whether any improvement can be observed over time. Only the highest scoring D1 and D4 showed some marginal development in quality over the time scale investigated (Table 1). GLs seem to have become more specific in stating their objectives (I1) and in creating more focused clinical questions (I2), and the recommendations in GLs have become more easily identifiable (I17) (online Data Supplement 2). However, the poor performance in D6 showed further deterioration from

Table 1. Critical appraisal of diabetes mellitus guidelines by the AGREE instrument.

Guideline	Date of issue	Source[a]	Scope[b]	Length, pages	GL manual[c]	Origin	Domain score, %						Overall assessment
							D1[d]	D2	D3	D4	D5	D6	
(12)	1999	Both	Diagnostic	>100	No	USA	89	40	69	35	17	21	Recommend with alteration
(13)	2001	Database	Combined	51–100	Yes	UK	56	75	74	71	8	71	Strongly recommend
(14)	2002	Journal	Combined	1–10	No	North America	47	6	17	33	14	4	Would not recommend
(15)	2002	Both	Diagnostic	11–50	No	USA	53	23	31	67	11	17	Recommend with alteration
(16)	2002	Database	Combined	>100	Yes	UK	92	85	87	98	33	42	Strongly recommend
(17)	2002	Database	Combined	>100	Yes	UK	92	88	90	98	33	42	Strongly recommend
(18)	2002	Database	Combined	1–10	No	South Africa	14	23	6	56	0	0	Would not recommend
(19)	2002	Both	Diagnostic	1–10	No	Canada	81	21	40	73	19	13	Recommend with alteration
(20)	2003	Database	Combined	>100	No	Canada	86	33	60	90	25	42	Recommend with alteration
(21)	2003	Database	Combined	>100	Yes	New Zealand	86	83	76	96	56	100	Strongly recommend
(22)	2003	Database	Diagnostic	51–100	Yes	USA	97	21	77	90	39	88	Strongly recommend
(23)	2003	Database	Diagnostic	>100	Yes	USA	94	23	74	81	42	83	Strongly recommend
(24)	2003	Database	Diagnostic	51–100	Yes	International	100	33	29	52	58	42	Recommend with alteration
(25)	2004	Both	Combined	1–10	Yes	USA	97	27	64	79	6	75	Recommend with alteration
(26)	2004	Database	Combined	1–10	No	USA, Canada	42	27	6	65	0	0	Would not recommend
(27)	2004	Database	Combined	>100	Yes	UK	97	88	92	98	72	92	Strongly recommend
(28)	2005	Database	Diagnostic	11–50	Yes	Canada	72	35	13	85	69	29	Recommend with alteration
(29)	2005	Database	Combined	51–100	Yes	International	58	46	55	79	44	96	Recommend with alteration
(30)	2005	Database	Diagnostic	>100	Yes	Australia	97	73	90	81	39	21	Strongly recommend
(31)	2006	Database	Diagnostic	51–100	Yes	International	78	15	26	69	25	21	Wouldn't recommend
(32)	2007	Both	Combined	>100	Yes	USA	64	42	39	69	17	46	Recommend with alteration
(33)	2007	Both	Combined	11–50	Yes	USA	61	31	39	92	39	0	Recommend with alteration
(34)	2007	Database	Diagnostic	0–50	Yes	International	86	27	55	60	28	9	Recommend with alteration
(35)	2007	Database	Combined	51–100	Yes	UK	97	71	64	90	56	21	Strongly recommend
(36)	2007	Database	Combined	51–100	Yes	UK	97	69	67	88	56	21	Strongly recommend
(37)	2007	Database	Combined	51–100	Yes	UK	75	71	67	81	72	29	Strongly recommend
Mean							77	45	54	76	34	39	
Range							14–100	6–88	6–92	33–98	0–72	0–100	

[a] Both, journal and electronic guideline database.
[b] Combined, diagnostic and therapeutic recommendations.
[c] GL development manual or technical document was available before GL publication.
[d] D1, scope and purpose; D2, stakeholder involvement; D3, rigor of development; D4, clarity and presentation; D5, applicability; D6, editorial independence.

Table 2. Performance of AGREE item scores in diabetes mellitus guidelines.

Domain and item		All GLs (n = 26)		
		Mean score	Score ≥3, n	Rate with score ≥3, %
D1[a]				
1	The overall objective of the guideline is specifically described.	3.34	19	73
2	The clinical questions covered by the guideline are specifically described.	3.16	17	65
3	The patients to whom the guideline is meant to apply are specifically described.	3.45	21	81
D2				
4	The guideline development team involves all relevant professional groups.	2.53	9	35
5	The patients' views and preferences have been sought.	2.12	9	35
6	The target users of the guideline are clearly defined.	3.03	15	57
7	The guideline has been piloted among target users.	1.75	6	23
D3				
8	Systematic methods were used to search for evidence.	2.50	10	38
9	The criteria for selecting the evidence are clearly described.	2.42	11	42
10	The methods used for formulating the recommendations are clearly defined.	2.30	7	27
11	The health benefits, side effects, and risks have been considered.	3.14	17	65
12	There is an explicit link between the recommendations and the supporting evidence.	3.12	19	73
13	The guideline has been externally reviewed by experts before its publication.	2.58	11	42
14	A procedure for updating the guideline is provided.	2.31	10	38
D4				
15	The recommendations are specific and unambiguous.	3.57	22	85
16	The different options for management of the condition are clearly presented.	3.37	21	81
17	The recommendations are easily identifiable.	3.75	24	93
18	The guideline is supported with tools for application.	2.43	10	38
D5				
19	Potential barriers in applying the recommendations have been discussed.	1.88	4	15
20	Potential cost implications of applying the recommendations have been considered.	2.21	6	23
21	The guideline presents key review criteria for monitoring and/or audit purposes.	1.95	8	31
D6				
22	The guideline is editorially independent from the funding body.	2.60	11	42
23	Conflicts of interest of guideline development members have been recorded.	1.95	8	31

[a] D1, scope and purpose; D2, stakeholder involvement; D3, rigor of development; D4, clarity and presentation; D5, applicability; D6, editorial independence.

2005 onward, with failures to report editorial independence and conflict of interest in the majority of GLs (Table 1).

Type of publication. We investigated how GL developers defined the type of their publications and whether these reflected the methods used for their develop-

ment. There was diversity in definitions: 19 publications were labeled as GLs or recommendations, of which 7 stated that they were evidence-based, 4 position statements or reports, and 3 guidance documents (Table 3). Among the 7 evidence-based GLs, 5 had evidence summaries and 6 graded recommendations. Three GLs that had evidence tables did not define their publications as evidence-based GLs *(22, 23, 25)*. More than two-thirds of GLs (n = 18, 69%) defined their grading system, but only 16 (62%) graded their final recommendations (Table 3).

Procedure for updating guidelines. Item 14 investigates whether GL developers describe the procedures for updating recommendations, including the timescale, responsibilities, and methods used. Fifteen GLs (58%) gave a timescale or expiration date, of which 1 GL provided this information in a separate GL development manual of the issuing authority (Table 3). The most frequent review date was 3 and 4 years. Only 10 GLs (38%) provided adequate information on the updating process (Table 2).

Reporting of laboratory-specific information in diagnostic guidelines. We investigated whether GLs covered essential laboratory-specific information, such as prevalence/pretest probability and diagnostic accuracy data or preanalytical and analytical factors critical for the correct interpretation and application of laboratory results in clinical practice (Table 3). About 60% of the GLs mentioned these factors. Reporting these pieces of information was more frequent in diagnostic compared to combined GLs, but the difference was not statistically significant in the various GL subgroups, as discussed below (online Data Supplement 3).

SUBGROUP ANALYSIS

GLs were grouped according to source, scope, length, origin, and availability of a guideline methods manual, to investigate whether there were statistically significant differences in GL quality in these subsets. Results are shown in Table 4 and online Data Supplements 3 and 4.

Subgrouping by source. Grouping GLs by source of publication revealed that 1 GL was published in a peer-reviewed journal, 19 were available in electronic GL databases, and 6 in both sources. The GL that was published exclusively in a peer-reviewed journal *(14)* was not recommended for use by the assessors. None of the 6 GLs published in both peer-reviewed journals and GL databases were strongly recommended. GLs published in electronic guideline databases only received a more favorable overall assessment. A notable difference, at a level of significance of $P \leq 0.05$, could be observed in the D5 only for the electronic GLs (Table 4).

Subgrouping by scope. The rate of occurrence of strongly recommended GLs was higher for the combined (50%) than for the diagnostic (30%) GLs, but the rate of GLs not recommended was also higher in the combined group. The difference was moderate ($P \leq 0.05$) in D2 only, with combined GLs scoring higher (Table 4). Moderate differences were also found in 4 individual items (online Data Supplement 4). Diagnostic GLs defined their objectives better (I1) and considered the cost implications of the recommendations more frequently (I20), whereas combined GLs defined their target users (I6) and their updating processes more precisely (I14) than diagnostic ones.

Subgrouping by length. A clear relationship could be demonstrated between GL length and methodological quality (Table 4). Most GLs that were not recommended were shorter, and all strongly recommended guidelines were longer than 50 pages. Significant differences between these subgroups could be found for most domains, with higher quality of the longer GLs. Moderate differences ($P \leq 0.05$) could be observed with the "applicability" and "clarity and presentation" domains. However, the best-performing GLs, scoring >50% in the "applicability" domain *(21, 24, 27, 28, 35–37)*, were generally longer than 50 pages, and all were published in electronic databases (Table 1).

Subgrouping by origin. The majority of the strongly recommended GLs (7 of 11) originated from the UK; the other 4 were from New Zealand, Australia, and the USA (Table 1). Significant differences ($P \leq 0.01$) could be observed in fulfilling the criteria of D2, with higher scores for the British GLs. In the "rigor of development" and "clarity and presentation" domains, the difference was moderate ($P \leq 0.05$) (Table 4).

Subgrouping by availability of guideline methods manual. Two-thirds of GLs had some accompanying manuals describing the methods of their development in some form. All strongly recommended GLs had such a manual (Table 4). All domain scores were better in the subset where these manuals were available, and the differences were highly statistically significant ($P \leq 0.01$) in D4, D5, and D6. In D1, D2, and D3, the *P* values were also significant, but at values somewhat >0.01.

Discussion

In the current study, we made special efforts to retrieve and use all available background technical materials when appraising GLs to avoid a biased assessment of methodological quality by the AGREE Instrument. Our evaluation revealed that diagnostic recommendations in the field of DM suffer from the same methodological weaknesses as GLs developed by prestigious

Table 3. Qualitative analysis of reporting diabetes mellitus guidelines.

Guideline	Type of publication as described by authors	Evidence table	Description of grading system	Graded recommendations	Review date, year	Prevalence/Pretest probability	Diagnostic accuracy	Preanalytical information	Analytical information
(12)	Review of the evidence and recommendations	+	–	–	–	–	+	–	–
(13)	National clinical guidelines	–	+	+	3	+	+	–	+
(14)	Consensus report	–	–	–	1	–	–	–	–
(15)	Guidelines and recommendations	–	+	+	–	+	+	+	+
(16)	Clinical guidelines and evidence review	+	+	+	4	–	+	–	+
(17)	Clinical guidelines and evidence review	+	+	+	4	–	–	–	–
(18)	Guideline	–	–	–	–	–	–	–	–
(19)	Clinical practice guidelines	–	+	+	–	+	+	+	+
(20)	Clinical practice guidelines	–	+	+	–	+	+	+	+
(21)	Evidence-based best practice guidelines	–	+	+	3	+	–	+	+
(22)	Recommendation and rationale statement	+	+	+	–	+	+	+	+
(23)	Recommendation and rationale statement	+	+	+	–	+	+	+	+
(24)	Report	–	–	–	–	+	+	–	–
(25)	Clinical practice guidelines	+	–	–	5	–	–	–	–
(26)	Guidelines	–	–	–	–	–	–	–	–
(27)	Clinical guidelines and evidence review	+	+	+	4	–	+	+	+
(28)	Guidelines and protocols	–	–	–	3	–	–	+	–
(29)	Global guideline	–	+	–	3–5	+	+	+	+
(30)	Evidence-based guidelines	+	+	+	3	+	+	+	+
(31)	Report	–	–	–	–	+	+	+	–
(32)	Medical guidelines (evidence based)	–	+	+	–	+	–	+	–
(33)	Position statement	–	+	+	1[a]	+	+	+	+
(34)	Guideline	–	+	+	3	–	–	–	–
(35)	Guidance	–	+	+	continuous	–	–	+	+
(36)	Guidance	–	+	–	continuous	+	+	+	+
(37)	Guidance	–	+	+	continuous	+	–	+	+
Percentage of GLs fulfilling criteria		31	69	62	58	58	58	62	58

[a] Information on updating is provided in a separate guideline development manual.

Table 4. Subgroup analysis

	D1[a]			D2			D3			D4			D5			D6					
	Mean domain score, %	SE	Range	Mean domain score, %	SE	Range	Mean domain score, %	SE	Range	Mean domain score, %	SE	Range	Mean domain score, %	SE	Range	Mean domain score, %	SE	Range	Strongly recommended, n (%)	Recommended with alteration, n (%)	Wouldn't recommend, n (%)
Source																					
Guideline database (n = 19)	80	5.2	14–100	52	6.1	15–88	58	6.5	6–92	80	3.3	52–98	40	5.1	0–72	45	7.6	0–100	11 (58)	5 (26)	3 (16)
Journal and GL database (n = 7)[b]	70	7.1	47–97	27	4.6	6–42	43	6.8	17–69	64	8.3	33–92	18	3.9	6–39	25	10.0	0–75	0 (0)	6 (86)	1 (14)
P	0.209			0.055			0.169			0.083			0.018[c]			0.152					
Scope																					
Diagnostic (n = 10)	85	4.5	53–100	31	5.2	15–73	50	8.2	13–90	69	5.3	35–90	35	5.8	11–69	34	8.9	9–88	3 (30)	6 (60)	1 (10)
Combined (n = 16)	73	6.2	14–97	54	6.8	6–88	56	6.9	6–92	80	4.5	33–98	33	6.1	0–72	43	8.8	0–100	8 (50)	5 (31)	3 (19)
P	0.286			0.023[c]			0.660			0.097			0.776			0.551					
Length																					
1–50 pages (n = 9)	61	8.5	14–97	24	2.7	6–35	30	7.0	6–64	68	5.8	33–92	21	7.4	0–69	16	8.0	0–75	0 (0)	6 (67)	3 (33)
>50 pages (n = 17)	86	3.5	56–100	56	6.2	15–88	67	4.8	26–92	80	4.1	35–98	41	4.6	8–72	52	7.2	21–100	11 (65)	5 (29)	1 (6)
P	0.009[d]			0.003[d]			0.001[d]			0.051			0.018[c]			0.001[d]					
Origin																					
North America (n = 12)	74	5.7	42–97	27	2.8	6–42	44	7.1	6–77	72	5.7	33–92	25	5.5	0–69	35	9.3	0–88	2 (17)	8 (66)	2 (17)
British (n = 7)	87	5.9	56–97	62	3.2	69–88	67	4.5	64–92	82	3.8	71–98	42	8.9	0–72	44	10.1	0–92	7 (100)	0 (0)	0 (0)
Other (n = 7)	74	11.3	14–100	43	9.8	15–83	48	11.2	6–90	70	5.9	52–96	36	7.6	0–58	41	15.4	0–100	2 (28.5)	3 (43)	2 (28.5)
P	0.355			0.001[d]			0.028[c]			0.037[c]			0.112			0.606					
Manual																					
yes (n = 19)	84	3.5	56–100	53	5.9	15–88	62	5.3	52–98	82	3.0	50–98	42	4.5	6–72	49	7.4	0–100	11 (58)	7 (37)	1 (5)
No (n = 7)	59	10.5	14–89	25	4.0	6–40	33	9.5	6–69	60	7.7	33–90	12	3.6	0–25	14	5.6	0–42	0 (0)	4 (57)	3 (43)
P	0.013[c]			0.015[c]			0.022[c]			0.010[d]			0.001[d]			0.004[d]					

[a] D1, scope and purpose; D2, stakeholder involvement; D3, rigor of development; D4, clarity and presentation; D5, applicability; D6, editorial independence.
[b] One guideline *(14)* was published in journal only.
[c] $P \leq 0.05$.
[d] $P \leq 0.01$.

authorities in many other disciplines *(2, 5, 38)*. Subgroup analyses of our study demonstrated that longer and electronically published GLs and the availability of GL development manuals yielded higher methodological scores in most AGREE domains (Table 4). One simple explanation is the lack of space available in journals for detailed and accurate reporting *(39)*. Poor methodological scores could just as well reflect faulty methods that could lead to biased and/or inaccurate conclusions. Paradoxically, lengthy GLs are thought to be less practical for daily use *(39)*, so one may argue that length of GLs adversely affects implementation. In our case, GLs that achieved high scores for "applicability" were indeed longer documents, but they also covered additional information on organization, cost implications, and monitoring of the use of recommendations in practice. All these tools help GL implementation, and thus we cannot confirm that lengthy GLs are not applicable in practice. The Conference on Guideline Standardization defined a standard for GL reporting to promote quality and facilitate implementation *(11)*. Such GL reporting standards have not yet been adopted by most journals, and peer reviewers also rarely use the AGREE or other criteria for systematic assessment of recommendations before publication *(40–42)*. These shortcomings suggest the need for GL reporting standards, similar to the STARD document for reporting diagnostic accuracy studies *(43)*, and clear publication and peer review policies for GLs by major medical journals.

In our study, the quality of purely diagnostic GLs was not significantly different from that of combined GLs (Table 4). Our additional evaluation in Table 3 showed that nearly half of all GLs do not report preanalytical, analytical, and diagnostic accuracy data *(3)*, which may lead to inappropriate interpretation of test results in clinical practice *(44)*. Fulfilling these criteria would be desirable in any GLs that deal with laboratory testing–related issues, since it is expected that practice recommendations are developed in a multidisciplinary process *(45)*. Unfortunately, this could not be confirmed by our study, as only 41% of the criteria were fulfilled in D2, which explored the involvement of all relevant stakeholders in the GL development process.

All GLs that scored better in the comparison by origin were from agencies with detailed GL manuals that provided a clear description and standards for the development process (Table 1). The availability of a GL manual, however, does not always guarantee that GL teams follow those processes consistently, and it has been shown that it is often not clear how decisions are made by the GL team when arriving at final recommendations *(8)*. The substantial heterogeneity, in both how the type of publication is defined and the adherence to this definition in the final presentation of the GL, suggests that there is likely a disparity between the methodology GL developers describe and what is actually followed in practice (Table 3). We found several GLs that described a grading system but did not grade their final recommendations. The lack of evidence tables in GLs that claim to be evidence-based may also point to potential deviations from the processes set in GL manuals. Therefore it is advisable that diagnostic GL development teams adhere to preset methodology and document the procedures followed explicitly.

We could not demonstrate major improvements in GL quality for most domains, and in the "editorial independence" domain, deterioration in scores was observed over time. We further evaluated the quality of GLs over time in some cases where the authorities issued several GLs [e.g., National Institute for Clinical Excellence (NICE), WHO, International Diabetes Federation (IDF)] within the time scale investigated (data not shown). The NICE GL in 2004 is of higher quality than the NICE 2002 version due to improvements in "applicability" and "editorial independence" domains. It is noteworthy that many international organizations have improved the rigor of their guideline development process and are moving toward international standardization *(11, 46–48)*. Surprisingly, the international WHO and IDF GLs in 2006 and 2007 had lower scores in most domains than the 2003 and 2005 versions, despite the fact that both agencies released guideline development manuals in 2003 (http://whqlibdoc.who.int/hq/2003/EIP_GPE_EQC_2003_1.pdf, http://www.idf.org). Therefore, we assume that the lower AGREE scores are due to the lack of reporting some methodological details rather than the lack of following the methodology described in the manuals. Explicit reporting of methodology and adherence to that methodology is particularly important for influential agencies (e.g., American Diabetes Association and WHO) whose recommendations are followed or adapted worldwide.

There are several limitations in our study. By evaluating English publications only, our results may suffer from language bias. However, several publications, including our own review of the topic, confirm no significant differences in the quality of English vs non-English publications of guidelines or trials *(2, 49, 50)*. Because most national DM GLs are based on or strongly influenced by international recommendations primarily published in English, we believe our results are likely to be generalizable.

Our study evaluated different publications that were defined in various ways by their authors. Such heterogeneity of definitions (such as guideline, guidance, protocols, position statement, recommendation and rationale statement, consensus report) may highlight different approaches in formulating recommen-

dations for practice. We also found several GLs that, while having proof of using evidence-based methods, failed to define their publication as such *(22, 23, 25)*. This suggests that the definitions used in the international guideline community may be confusing for both GL developers and users, and that simplification and standardization of terminology is needed. One may argue that AGREE can be used for assessing evidence-based GLs only. However, AGREE is a generic and widely accepted toolbox *(8)* that can investigate the GL development process irrespective of whether it applies evidence- or consensus-based methodology *(4)*. In fact, most evidence-based GLs have a substantial element of consensus-based judgment, especially when evidence is conflicting or lacking. In the latter case, GL developers should still search for and appraise the "best available" evidence before they conclude that the best they can do is to reach consensus.

Our study does not determine whether there are relationships between the methodological quality of GLs and the validity of their content. The AGREE Instrument or other GL appraisal tools can investigate neither the accuracy of the content of recommendations nor their impact on patient outcomes *(51, 52)*. Another shortcoming of all critical appraisal tools is that they do not differentiate between whether the publication fails certain criteria due to lack of reporting or to poor methodology and design. Therefore, our results should not be interpreted as criticisms of the truth of scientific statements or the validity of recommendations made in a given publication. However, the demonstrated shortcomings in reporting and/or the methodology applied by different GL developers could lead to distrust in and/or misuse of recommendations *(53)*. With such shortcomings, the energy put into developing scientifically accurate but otherwise poorly presented GLs could end up being wasted, whereas inaccurate but otherwise nicely presented GLs might be promoted and used widely. This is why we advise that GLs be critically evaluated for both methodology and content before recommendations are used in clinical practice *(38)*.

In conclusion, our results suggest the need for systematically developed, explicit recommendations based on evidence-based guideline development and reporting standards in laboratory medicine. Our study also highlights the need for simplification and standardization of GL terminology. Further studies are needed to explore in depth the relationship between the scientific validity and the methodological quality of diagnostic recommendations in DM.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and has met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: Several authors of this article (J. Watine, W. Oosterhuis, D. Rogic, S. Sandberg, P. S. Bunting, A.R. Horvath) are members of the Committee on Evidence-Based Laboratory Medicine of IFCC, and this work was carried out in collaboration with the IFCC Task Force on the Global Campaign for Diabetes Mellitus.

References

1. Field MJ, Lohr KN, eds. Guidelines for clinical practice: from development to use. Washington, DC: National Academy Press; 1992. 426 p.
2. Horvath AR, Nagy E, Watine J. Critical appraisal of guidelines. In: Evidence-based Laboratory Medicine: Principles, Practice, Outcomes. Price CP, Christenson RH (eds) AACC Press, Washington. 2nd edition, 2007;295–319.
3. Oosterhuis WP, Bruns DE, Watine J, Sandberg S, Horvath AR. Evidence-based guidelines in laboratory medicine: principles and methods. Clin Chem 2004;50:806–18.
4. The AGREE Collaboration. Appraisal of Guidelines for Research and Evaluation (AGREE) Instrument. http://www.agreecollaboration.org (Accessed December 2007).
5. Horvath AR, Nagy E, Watine J. Quality of guidelines for the laboratory management of diabetes mellitus. Scand J Clin Lab Invest Suppl 2005;240: 41–50.
6. Hunt DL, McKibbon KA. Locating and appraising systematic reviews. Ann Intern Med 1997;126: 532–8.
7. Horvath AR, Pewsner D. Systematic reviews in laboratory medicine: principles, processes and practical considerations. Clin Chim Acta 2004; 342:23–39.
8. Van der Wees PJV, Hendriks EJM, Custers JWH, Burgers JS, Dekker J, de Bie RA. Comparison of international guideline programs to evaluate and update the Dutch program for clinical guideline development in physical therapy. BMC Health Services Research 2007;7:191. http://www.biomedcentral.com/1472–6963/7/191 (Accessed December 2007).
9. The AGREE Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical guidelines: the AGREE project. Qual Saf Health Care 2003;12: 18–23.
10. MacDermid JC, Brooks D, Solway S, Switzer-McIntyre S, Brosseau L, Graham ID. Reliability and validity of the AGREE instrument used by physical therapists in assessment of clinical practice guidelines. BMC Health Services Research 2005;5:18. (http://www.biomedcentral.com/1472–6963/5/18) (Accessed December 2007).
11. Shiffman RN, Shekelle P, Overhage JM, Slutsky J, Grimshaw J, Deshpande AM. Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization. Ann Intern Med 2003;139:493–8.
12. Woolf SH, Davidson MB, Greenfield S, Bell HS, Ganiats TG, Hagen MD et al. American Academy of Family Physicians and American Diabetes Association. The benefits and risk of controlling blood glucose levels in patients with type 2 diabetes mellitus: a review of the evidence and recommendations. AAFP Policy Action April 1999. http://www.aafp.org/online/etc/medialib/aafp_org/documents/clinical/clin_recs/diabetespolicy.Par.0001.File.tmp/clinicalrecs_diabetespolicy05.pdf (Accessed December 2007).
13. Scottish Intercollegiate Guidelines Network. Management of diabetes CPG55. 2001. http://www.sign.ac.uk/guidelines/fulltext/55/index.html

(Accessed December 2007).
14. Reece EA, Homko C, Miodovnik M, Langer O. A consensus report of the Diabetes in Pregnancy Study Group of North America Conference, Little Rock, Arkansas, May 2002. J Matern Fetal Neonatal Med 2002;12:362–4.
15. Sacks DB, Bruns DE, Goldstein DE, Maclaren NK, McDonald JM, Parrott M. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. Clin Chem 2002;48:436–72.
16. McIntosh A, Hutchinson A, Home PD, Brown F, Bruce A, Damerell A, Davis R et al. (2001) Clinical guidelines and evidence review for type 2 diabetes: management of blood glucose. Sheffield: ScHARR, University of Sheffield. http://www.nice.org.uk/cat.asp?c=36733 (Accessed December 2007).
17. McIntosh A, Hutchinson A, Feder G, Durrington P, Elkeles R, Hitman GA, et al. (2002) Clinical guidelines and evidence review for type 2 diabetes: lipids management. Sheffield: ScHARR, University of Sheffield. http://www.nice.org.uk/cat.asp?c=38551 (Accessed December 2007).
18. Society for Endocrinology, Metabolism and Diabetes of South Africa. Revised SEMDSA Guidelines for diagnosis and management of type 2 diabetes mellitus for primary health care in 2002. http://www.semdsa.org.za/guidelines.htm (Accessed December 2007).
19. Berger H, Crane J, Farine D, Armson A, De La Ronde S, Keenan-Lindsay L, et al. Screening for gestational diabetes mellitus. J Obstet Gynaecol Can 2002;24:894–912.
20. Canadian Diabetes Association Clinical Practice Guidelines Expert Committee. Canadian Diabetes Association 2003 Clinical Practice Guidelines for the Prevention and Management of Diabetes in Canada. http://www.diabetes.ca/cpg2003/downloads/cpgcomplete.pdf (Accessed December 2007).
21. New-Zealand Guidelines Group. Management of Type 2 Diabetes. http://www.nzgg.org.nz/guidelines/dsp_guideline_popup.cfm?guidelineCatID=32&guidelineID=36 (Accessed December 2007).
22. U.S. Preventive Services Task Force. Screening for type 2 diabetes mellitus in adults: recommendations and rationale. February 2003. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/clinic/uspstf/uspsdiab.htm (Accessed December 2007).
23. U.S. Preventive Services Task Force. Screening for gestational diabetes mellitus: recommendations and rationale. http://www.ahrq.gov/clinic/uspstf/uspsgdm.htm (Accessed December 2007).
24. World Health Organization. Screening for type 2 diabetes: report of a World Health Organization and International Diabetes Federation meeting. Geneva, 2003. http://whqlibdoc.who.int/hq/2003/WHO_NMH_MNC_03.1.pdf (Accessed December 2007).
25. Snow V, Aronson MD, Hornbake ER, Mottur-Pilson C, Weiss KB. Clinical Efficacy Assessment Subcommittee of the American College of Physicians. Lipid control in the management of type 2 diabetes mellitus: a clinical practice guideline from the American College of Physicians. Ann Intern Med 2004;140:644–9.
26. Kaiser Permanente, Care Management Institute's Diabetes Guidelines workgroup. Guidelines for the Management of Adult Diabetes in Primary Care. http://members.kaiserpermanente.org/kpweb/pdf/feature/247clinicalpracguide/CMI_AdultDiabetesGuideline_public_web_033104.pdf (Accessed December 2005).
27. National Collaborating Centre for Women's and Children's Health (NICE guidance). Type 1 diabetes: diagnosis and management of type 1 diabetes in children and young people. http://www.nice.org.uk/page.aspx?o=213575 (Accessed December 2007).
28. Guidelines and Protocols Advisory Committee (Br Columbia Ministry of Health). Diabetes Care, revised 2005. http://www.health.gov.bc.ca/gpac/guideline_diabetes.html (Accessed December 2007).
29. IDF Clinical Guidelines Task Force. Global guideline for type 2 diabetes. Brussels: International Diabetes Federation, 2005. http://www.idf.org/home/index.cfm?node=1457 (Accessed December 2007).
30. National Health and Medical Research Council. National evidence based guidelines for the management of type 2 diabetes mellitus. http://www.health.gov.au/nhmrc/publications/pdf/cp86.pdf (Accessed November 2007).
31. World Health Organization: Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF Consultation. Geneva, 2006. http://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf (Accessed December 2007).
32. American Association of Clinical Endocrinologists. American Association of Clinical Endocrinologists medical guidelines for clinical practice for the management of diabetes mellitus. Endocr Pract 2007;13(Suppl 1):1–68.
33. American Diabetes Association. Standards of medical care in diabetes. Diabetes Care 2007;30:S4–41.
34. IDF Clinical Guidelines Task Force. Guideline for management of postmeal glucose. Brussels: International Diabetes Federation, 2007. http://www.idf.org/home/index.cfm? node=1620 (Accessed December 2007).
35. Sowerby Centre for Health Informatics at Newcastle. PRODIGY Guidance: Diabetes type 2: lipid management. http://www.prodigy.nhs.uk/guidance.asp?gt=Diabetes%20-%20lipid%20management (Accessed December 2007).
36. Sowerby Centre for Health Informatics at Newcastle: PRODIGY Guidance: Diabetes type 2: renal disease. http://www.prodigy.nhs.uk/guidance.asp?gt=Diabetes%20-%20renal%20disease (Accessed December 2007).
37. Sowerby Centre for Health Informatics at Newcastle: PRODIGY Guidance: Blood glucose. http://www.prodigy.nhs.uk/guidance.asp?gt=Diabetes%20-%20renal%20disease (Accessed December 2007).
38. Qaseem A, Vijan S, Snow V, Cross T, Weiss KB, Owens DK, for the Clinical Efficacy Assessment Subcommittee of the American College of Physicians. Glycaemic control and type 2 diabetes mellitus: the optimal hemoglobin A_{1c} targets. A guidance statement from the American College of Physicians. Ann Intern Med 2007;147:417–22.
39. Deeks JJ. Word limits best explain failings of industry supported meta-analyses [Letter]. BMJ 2006;333:1021.
40. Fervers B, Burgers JS, Haugh MC, Brouwers M, Browman G, Cluzeau F, Philip T. Predictors of high quality clinical practice guidelines: examples in oncology. Int J Qual Health Care 2005;17:123–32.
41. Miller J, Petrie J. Development of practice guidelines [Commentary]. Lancet 2000;355:82–3.
42. vanTulder MW, Tuut M, Pennick V, Bombardier C, Assendelft WJJ. Quality of primary care guidelines for acute low back pain. Spine 2004;29:E357–62.
43. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7–18.
44. Skeie S, Nordin G, Oosterhuis WP, Araczki A, Horvath AR, Perich C et al. Post-analytical external quality assurance of blood glucose and HbA_{1c}: an international survey. Clin Chem 2005;51:1145–53.
45. Grimshaw GM, Khunti K, Baker R. Diagnosis of heart failure in primary care: an assessment of international guidelines. Br J Gen Pract 2001;51:384–6.
46. Raine R, Sanderson C, Black N. Developing clinical guidelines: a challenge to current methods. BMJ 2005;331:631–3.
47. Oxman AD, Fretheim A, Schünemann HJ, SURE. Improving the use of research evidence in guideline development: introduction. Health Res Policy Syst 2006;4:12.
48. Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: I. Guideline for guidelines. Health Res Policy Syst 2006;4:13.
49. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, Le Lorier J, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet 1996;347:363–6.
50. Burgers JS, Grol R, Klazinga NS, Makela M, Zaat J. Towards evidence-based clinical practice: an international survey of 18 clinical guideline programs. Int J Qual Health Care 2003;15:31–45.
51. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers Dirk. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. Int J Qual Health Care 2005;17:235–42.
52. Burgers JS. Guidelines quality and guidelines content: are they related [Editorial]. Clin Chem 2006;52:3–4.
53. Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Potential benefits, limitations, and harms of clinical guidelines. BMJ 1999;318:527–30.