The basis for monitoring strategies in clinical guidelines: a case study of prostate-specific antigen for monitoring in prostate cancer

Jacqueline Dinnes PhD, Jenny Hewison PhD, Douglas G. Altman DSc, Jonathan J. Deeks PhD CStat

See related commentary by Johansson and Stattin on page 159 and at www.cmaj.ca/lookup/doi/10.1503/cmaj.111989

Abstract

Background: The volume of published literature on the evaluation and use of tests for monitoring purposes is sparse. Our aim was to determine the extent to which recommendations for monitoring prostate-specific antigen to detect recurrent prostate cancer consider key factors that should inform rule-based strategies for monitoring.

Methods: We reviewed the recommendations made in clinical guidelines for the repeated measurement of prostate-specific antigen in men who have received primary treatment for localized prostate cancer. We assessed the guidelines using the Appraisal of Guidelines for Research and Evaluation Framework.

Results: We identified guidelines and statements of best practice from nine organizations. We saw considerable inconsistency in recommendations for testing for prostate-specific antigen as a form of monitoring. Recommendations on when to test appeared to be almost exclusively determined using standard follow-up schedules rather than any scientific basis. Recommendations on when to take action were primarily based on consensus statements or retrospective case series. Eight of the nine guidelines acknowledged the potential presence of measurement variability, but they did not attempt to account for the effect of such variability on the interpretation of the results of tests for prostate-specific antigen. Many recommendations were made with few or no supporting references; however, a variety of papers were cited across guidelines. Of 48 papers cited, 29.1% (14/48) were reviews; the remaining 70.8% (34/48) of papers cited were primary studies.

Interpretation: A systematic approach to the development of monitoring schedules using prostate-specific antigen in guidelines for prostate cancer is lacking, due to inadequacies in the available evidence and its use.

Monitoring involves the scheduled, repeated use of a test or tests in an individual over time to make decisions about the management of a disease or condition. It is a central activity in the management of care, taking up a considerable part of the clinical workload and associated cost.[1] In contrast, the volume of published literature on the evaluation and use of tests for monitoring purposes is relatively sparse.

Mant and others have provided a framework for developing and evaluating a monitoring strategy with four main steps:[1–3] deciding whether to monitor, choosing a test, specifying and assessing the monitoring strategy to be used, and implementing the strategy. Underlying this framework is the key concept that the "signal" from the test, reflecting the status of the underlying condition, should be greater than the surrounding "noise," or measurement variability, that may affect the interpretation of the results.[2,3] If the noise is too high in relation to the signal, one's certainty in a given test result will be considerably reduced.

The repeated measurement of prostate-specific antigen among men who have received primary treatment for prostate cancer is an apparently successful example of a rule-based monitoring strategy. The levels of prostate-specific antigen following radical treatment vary. Recurrence of disease following radical prostatectomy is associated with the presence of prostate-specific antigen; following radical radiotherapy, it is associated with a rise in the level of prostate-specific antigen.[4] When a predefined level is reached, biochemical failure is said to have occurred. The usefulness of testing for prostate-specific antigen as a form of monitoring is based on the assumption that biochemical failure predates clinical failure within some clinically meaningful time frame. The decision to initi-

Competing interests: Jacqueline Dinnes has received grants from the National Institute for Health Research (NIHR). Jenny Hewison has received grants from the NIHR and Siemens; she is a member of the board for the NIHR Health Technology Assessment programme, the NIHR Programme Grants for Applied Research, and the Fetal Anomaly Screening programme; she is a consultant for the UK Consumers' Association's *Which?* magazine; and she has received reimbursement for travel expenses from the European Society for Human Reproduction and Embryology. Douglas Altman has received reimbursement for travel expenses from the NIHR. Jonathan Deeks has received grants from the NIHR and is a member of the board for the NIHR Health Technology Assessment programme. No other competing interests were declared.

This article has been peer reviewed.

Correspondence to: Jacqueline Dinnes, j.dinnes@bham.ac.uk

***CMAJ* 2012. DOI:10.1503/cmaj.110600**

ate treatment for recurrence, however, will depend on multiple factors rather than on a single value.[5]

We reviewed clinical guidelines for recommendations for monitoring prostate-specific antigen for the detection of recurrent prostate cancer to determine the extent to which the guidelines consider key factors that should inform rule-based strategies for monitoring. In particular, we assessed the degree of consistency between guidelines, the explicit consideration of factors important for specifying a strategy for monitoring, and the use of supporting evidence to justify any recommendations.

Methods

Inclusion criteria

We included guidelines that considered a test for prostate-specific antigen for monitoring patients whose localized prostate cancer was treated with either radical prostatectomy or radical radiotherapy. We excluded guidelines that considered only screening or treatment. We did not consider recommendations for measuring prostate-specific antigen after other potentially curative treatments or as part of active surveillance.

Box 1: Framework criteria used to assess the rigour of guideline development*

- **Systematic methods of searching were used:** Details of the strategy used to search for evidence should be provided, including the search terms used, the sources consulted and the dates of the literature covered.
- **Selection criteria are clearly described:** Criteria for including or excluding evidence identified by the search should be provided. These criteria should be explicitly described, and the reasons for including or excluding evidence should be clearly stated.
- **Formulation of recommendations are clearly described:** There should be a description of the methods used to formulate the recommendations and how final decisions were made. Areas of disagreement and methods of resolving disagreement should be specified.
- **Recommendations should consider relevant issues for monitoring:†** For our study, these issues include the variability in measurements or the need for retesting, the rationale presented for the choice of testing interval and prostate-specific antigen threshold, and the acknowledgement of the uncertainties in the natural history of prostate-specific antigen following radical treatment.
- **Explicit link to supporting evidence:** There should be an explicit link between the recommendations and the evidence on which they are based. Each recommendation should be linked to a list of references on which it is based.
- **Guidelines underwent prepublication external review:** A guideline should be reviewed externally before it is published. A description of the method used for the external review should be presented, which may include a list of the reviewers and their affiliations.
- **Procedure for updating guidelines is described:** Guidelines need to reflect current research. There should be a clear statement about the procedure for updating the guidelines.

*This framework is adapted from the "Rigour of Development" section of the original Appraisal of Guidelines for Research and Evaluation instrument.[6]
†Original criterion related to the benefits and harms of interventions (i.e., "The guideline should consider health benefits, side effects, and risks of the recommendations.").

Literature searches

We conducted a Boolean search of MEDLINE from January 1999 to July 2009 for the following medical subject headings: ("Prostatic Neoplasms" or "Prostate-Specific Antigen") and "Practice Guideline." We limited our results to publications in English. We also searched the National Library of Guidelines, the Trip database, and the Cochrane Library and checked the reference lists of the papers retrieved for further relevant guidelines. The titles and abstracts of retrieved records were assessed independently for inclusion by Jacqueline Dinnes and Jonathan Deeks, who also resolved discrepancies by consensus.

Data extraction

We extracted all recommendations or statements relating to the use of tests for prostate-specific antigen following treatment with curative intent and noted references to any supporting evidence. We assessed the methods used to create the guidelines using the Appraisal of Guidelines for Research and Evaluation framework. This framework contains 23 key items organized into six domains.[6] We applied only the seven items included in the domain for "rigour of development" (Box 1; see Appendix 1, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.110600/-/DC1, for a full description of the criteria). We replaced the fourth item in this domain with one relevant to using tests for monitoring, as opposed to consideration of the benefits and harms of interventions.

We took a generous approach to scoring each of these items. For example, if a systematic search was reported to have been done but was not reported in detail, the guideline would score three out of a possible four points. If a discussion of evidence was provided that appeared to relate to a recommended monitoring schedule, an explicit link with evidence was judged to have been provided without closer examination of the actual evidence cited. We did not judge the acceptability of rationale presented for the frequency of testing or threshold values for results, but we did note whether or not a rationale was presented. A maximum score of four points was attached to each of the seven items, for a maximum total score of 28.

Synthesis

We conducted a narrative synthesis.

Results

We identified guidelines ($n = 7$) or statements on best practice ($n = 2$) from nine organizations,[7–15] four of which were from North America, four

from Europe and one from Australia. Nearly all of the guidelines scored poorly on the framework criteria, with scores ranging between 9 and 16 out of a possible 28 (Figure 1; see Appendix 1 for further details). The sole exception was the set of guidelines issued by the National Institute for Health and Clinical Excellence, which scored 22 points.[12] The highest scoring item of all of the framework criteria was the use of systematic searches. Such searches were reported in most of the guidelines we studied, although they were not often described in detail. Methods for the formulation of recommendations were well described (i.e., a description of the methods used, how final decisions were made and methods for resolving disagreement) in only three of the guidelines.[10–12] Only one guideline fully considered relevant issues for monitoring tests;[12] it was also the only one to consistently provide clear links between its recommendations and the underlying evidence base, and it reported the methods used in more detail than most of the other guidelines in our sample.

Table 1 shows the lack of consistency among guideline recommendations regarding the frequency of follow-up assessments and threshold values for results; there does not appear to be any clear pattern in recommendations over time.

Eight of the nine guidelines[7–14] acknowledged that levels of prostate-specific antigen may be affected by technical or biologic variability. In most cases, this information was presented in the introductory sections of the guidelines; only one set of guidelines tempered its recommendations with reference to a single measurement of prostate-specific antigen possibly being unreliable and recommended retesting within two months.[12]

Three guidelines acknowledged the potential impact of technical variation,[11–13] recommending that the same assay be used for each measurement. Four guidelines made some attempt to justify the interval between tests,[7,10,12,15] and three guidelines discussed relevant issues affecting the choice of threshold values.[7,11,12] Three guidelines stated that it was not possible to provide a recommendation on the most appropriate definition of biochemical failure.[8,9,11]

Only three of the nine guidelines commented on the difficulty of using prostate-specific antigen as a monitoring tool owing to the uncertainties in its behaviour following radical treatment for prostate cancer.[9,11,12] Two sets of guidelines clearly recognized that not all men with biochemical failure go on to experience clinical failure such that evidence of the former alone may not be sufficient to alter treatment.[11,12]

Many recommendations on the frequency of testing or threshold levels for prostate-specific antigen were made with no or few supporting citations (Table 1). Only one guideline[10] cited a primary study in support of its recommended intervals for monitoring, and only four of the guidelines[10,12,14,15] showed the level of evidence supporting their recommendations. The levels of evidence cited ranged from consensus of the "Guideline Development Group" to "well-conducted clinical studies" (Table 1), suggesting that different groups had different views on the quality of the evidence available.

Despite the general lack of citations in individual guideline documents, a wide range of papers were cited across guidelines. A total of 48 papers were cited (Table 2 and Appendix 1); 29.1% (14/48) were reviews, and the remainder were primary studies, almost exclusively retrospective in nature. Of the 34 primary studies, we judged half (17) to have studied the natural history of prostate-specific antigen following treatment and eight to have evaluated the effect of different definitions of biochemical failure on clinical outcomes. Only two primary studies examining variability in measurements were cited.

Most of the studies were cited by only one or two of the guidelines; those references cited by three or more guidelines are presented in Table 3. Among the most frequently cited studies were two consensus statements[16,17] and a review

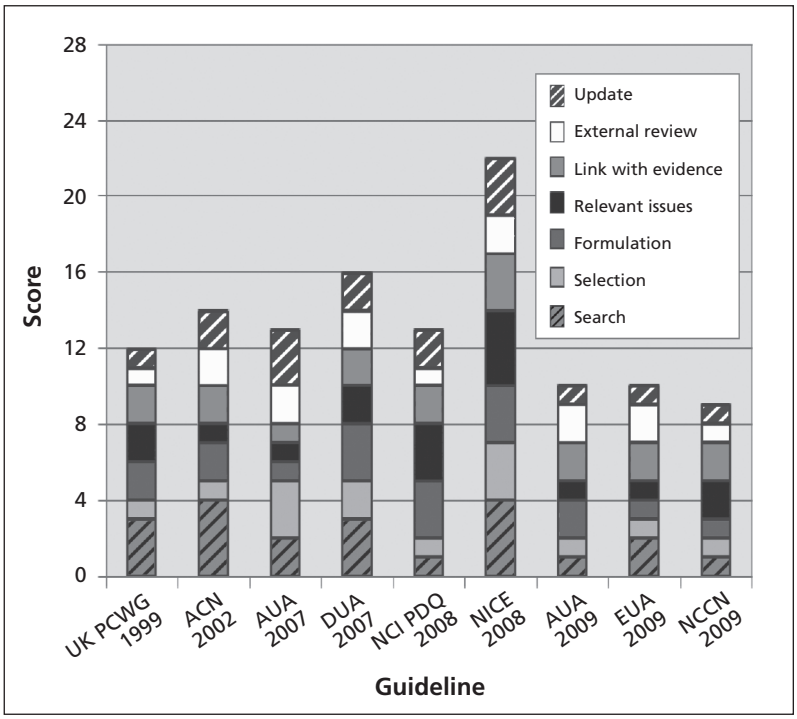

Figure 1: Chart showing the scores guidelines received for each of seven framework criteria used to assess the rigour of guideline development. UK PCWG = UK Prostate Cancer Working Group, ACN = Australian Cancer Network, AUA = American Urological Association, DUA = Dutch Urological Association, NCI PDQ = National Cancer Institute – Physician Data Query (US), NICE = National Institute for Health and Clinical Excellence (UK), EAU = European Association of Urology, NCCN = National Comprehensive Cancer Network (US).

Table 1: Statements or recommendations made in guidelines for monitoring with prostate-specific antigen from nine difference organizations and the types of supporting evidence cited (part 1 of 2)

Statement/recommendation	UK PCWG 1999[7]	ACN 2002[8]	AUA 2007[9]	DUA 2007[10]	NCI PDQ 2008[11]	NICE 2008[12]	AUA 2009[13]	EAU 2009[14]	NCCN 2009[15]	No. of guidelines
Frequency of follow-up visits after radical treatment										
Quarterly for 1* or 2† yrs, then every 6 mo or annually	+† ASTRO	–	–	++† 3 primary studies (level 3)‡	–	–	–	++* No supporting evidence cited§	–	3
Every 6 mos (for 2† or 5¶ yr), then annually	–	–	+¶ No supporting evidence cited	–	–	++† No direct evidence**	–	–	++‡ 1 primary study††	3
Threshold for intervention following prostatectomy										
Any detectable PSA	+ No supporting evidence cited	–	–	–	+ 3 primary studies	–	–	–	++ No supporting evidence cited††	3
PSA > 0.2 ng/mL	–	–	–	++ 1 primary study, 1 review (level 4)‡	–	+ 3 primary studies, 2 reviews	++ 1 primary study, 1 review	++ 4 primary studies, 2 reviews§	–	4
No definite threshold recommended	–	+ 1 primary study	+ No supporting evidence cited	–	–	–	–	–	–	2
Threshold for intervention following radiotherapy										
Three consecutive increases in PSA (ASTRO 1997 definition)[16]	+ ASTRO	–	–	++ ASTRO (level 4)‡	–	–	–	–	–	2
PSA nadir + 2 ng/mL (Phoenix 2005 definition)[17]	–	–	–	–	–	–	++ 2 primary studies, Phoenix	++ 2 primary studies, ASTRO, Phoenix§	++ Phoenix††	3

Table 1: Statements or recommendations made in guidelines for monitoring with prostate-specific antigen from nine difference organizations and the types of supporting evidence cited (part 2 of 2)

Statement/recommendation	UK PCWG 1999[7]	ACN 2002[8]	AUA 2007[9]	DUA 2007[10]	NCI PDQ 2008[11]	NICE 2008[12]	AUA 2009[13]	EAU 2009[14]	NCCN 2009[15]	No. of guidelines
PSA nadir + 4 ng/mL	–	–	–	–	–	+ 3 primary studies, Phoenix, 1 review	–	–	–	1
No specific recommendation	–	+ 2 primary studies, 1 review	+ No supporting evidence cited	–	+ 2 primary studies, ASTRO, Phoenix	–	–	–	–	3
Sources of PSA variability acknowledged and/or remedial action recommended										
Technical variability possible	+ No supporting evidence cited	+ 1 primary study	–	+ 1 review	+ 1 primary study	++**	+ 3 primary studies	+ 4 primary studies	–	7
Biologic variability possible	–	+ No supporting evidence cited	+ No supporting evidence cited	+ 3 primary studies	–	–	+ 18 primary studies	+ 8 primary studies	–	5
Remedial action recommended	–	–	–	++ repeat at 1–2 mo	+ Same assay	++ Same assay	+ Same assay, 3–6 wk postbiopsy	–	–	4
Acknowledgement of uncertainties in natural history of PSA and prostate cancer following primary treatment	–	–	++ 1 review	–	++ 4 primary studies	++ 1 primary study, 1 review	–	–	–	3

Note: + = factor was considered anywhere within the guideline document, ++ = factor was considered within the guideline recommendations, – = factor was not acknowledged in the document, ACN = Australian Cancer Network, ASTRO = American Society for Therapeutic Radiology and Oncology 1997 consensus statement,[16] AUA = American Urological Association, DUA = Dutch Urological Association, EAU = European Association of Urology, NCCN = National Comprehensive Cancer Network (US), NCI PDQ = National Cancer Institute – Physician Data Query (US), NICE = National Institute for Health and Clinical Excellence, PCWG = Prostate Cancer Working Group, Phoenix = 2005 revision of the ASTRO consensus statement,[17] PSA = prostate-specific antigen.
*Initial follow-up schedule to be followed for 1 yr.
†Initial follow-up schedule to be followed for 2 yr.
‡Guideline document reports levels of evidence as follows: level 3 corresponds to at least one randomized controlled trial or other comparative/noncomparative study; level 4 corresponds to expert opinion from people such as members of the working group.
§Guidelines document reports level of evidence as grade B, corresponding to well-conducted clinical trials, but without randomized controlled trials.
¶Initial follow-up schedule to be followed for 5 yr.
**Level of evidence reported in the guideline document corresponds to consensus of the Guidelines Development Group.
††Guidelines document reports level of evidence as level 2a, corresponding to lower-level evidence and uniform NCCN consensus.

of definitions of biochemical failure.[5] The four most frequently cited primary studies had the largest sample sizes of all of the primary studies cited; three of the four studies evaluated different definitions of biochemical failure,[18–20] and one studied the natural history of disease progression among men with elevated levels of prostate-specific antigen.[21]

Discussion

We found considerable inconsistency among the recommendations made in the different guidelines we studied in terms of using prostate-specific antigen as a monitoring tool, even when the guidelines were published within a few years of each other. Factors considered to be important when specifying a monitoring strategy were given limited attention and were not well-supported with reference to primary literature.

Recommendations on when to test and what action to take consequent to a given test result were very much considered in isolation from each other. "When to test" appeared to be almost exclusively determined by standard follow-up schedules rather than being based on any scientific evidence. Although most guidelines acknowledged the potential for variation in measurements, they did not attempt to account for the potential effect of such variation on the interpretation of test results. A systematic review of biologic variation in levels of prostate-specific antigen found mean variability of 20%.[22] A calculation of the reference change values suggested that to be 95% sure that a change in total prostate-specific antigen is not due to random variation, the change needs to be about 50% of the previous measurement.[22] This review was not cited by any of the seven guidelines subsequently published.

Recommendations on when to take action were based on consensus statements or retrospective case series with little attention paid to variations in the definition of the threshold, the definition of clinical failure, and the frequency and length of follow-up between studies; each of these factors can affect the accuracy of any given cut-off. Sensitivity and specificity are also known to be affected by differences in the case mixture between studies.[23] These differences were not acknowledged by any of the identified guidelines; however, a 2005 review of monitoring prostate cancer with prostate-specific antigen found it impossible to recommend any single definition of biochemical failure after treatment for this reason.[24] This review was cited by only one of the nine guidelines,[12] possibly because it

Table 2: Evidence used to support recommendations made in guidelines for monitoring with prostate-specific antigen

		Recommendation				
Type of study	No. of studies	Frequency of testing	Threshold (radical prostatectomy)	Threshold (radical radiotherapy)	Variability	Uncertainty in natural history of PSA
ASTRO consensus statements	3	x		x		
Best practice statement	1	x				
Reviews	10	x	x	x	x	x
Primary studies	34	x	x	x		x
Acceptability of follow-up	3	x				
Optimal frequency of follow-up	1	x				
Natural history of PSA following treatment	15	x	x	x		x
Natural history of PSA without treatment	1		x			
Salvage radical radiotherapy outcomes	4		x			
Testing biochemical failure definitions	8		x	x		x
Measurement variability	2				x	
No. of guidelines citing evidence		4	6	7	2	3

Note: ASTRO = American Society for Therapeutic Radiology and Oncology, PSA = prostate-specific antigen.

was not fully systematic. A relatively systematic method was used to identify studies for inclusion in the review, but most of the studies apparently meeting the inclusion criteria were excluded because of poor design, small patient numbers, insufficient follow-up, space limitations in journals or duplicate publication. The definitions of these criteria and the number of studies excluded for each reason were not clearly documented, therefore the review cannot be considered fully systematic. Given the lack of descriptions of inclusion criteria used in the guidelines, it is difficult to reconcile why an individual study or review was or was not included.

Reviews of guidelines in other areas have shown similar findings regarding the presentation of evidence for recommended schedules for monitoring.[25,26] Reviews of treatment[27] and diagnostic guidelines[28] have identified a similar inconsistency in recommendations between guidelines and variation in the evidence cited, with some referring to a substantial body of evidence and others presenting very little.[29–31]

Several factors are likely to contribute to our findings. First, although monitoring is starting to receive more attention,[32] there is a lack of high-quality evidence and clear guidance in terms of what to consider when establishing monitoring strategies. It is therefore perhaps not surprising that relevant evidence has not been used to inform guidelines.

Second, the various pieces of information needed to inform a monitoring strategy are not usually available from a single study. Ideally, one or more monitoring strategies should be evaluated in a randomized controlled trial or some form of prospective comparative study. Where there is high-quality evidence, greater consensus between guideline recommendations and stronger guideline recommendations have been found.[27] Randomized trials of monitoring, however, have their own challenges[33] and are consequently relatively rare. Instead, evidence has to be gathered from various sources. Although the diversity of evidence needed to inform coherent monitoring strategies makes the identification of relevant pieces of evidence a challenge for guideline developers and likely adds to the inconsistency in recommendations between guidelines, guideline developers have a responsibility to highlight recommendations where there is a lack of evidence or the evidence is inconsistent.

Efforts to improve the evidence base for monitoring are ongoing. For example, a Bayesian hierarchical changepoint model has been used to simulate the behaviour of prostate-specific anti-

Table 3: The seven most commonly cited studies in guidelines for screening with prostate-specific antigen from nine organizations

Study	Design/aim (as extracted from abstract)	Statement(s) supported by the study	No. of guidelines citing the study
Roach et al.[17] (2006)	Reports second consensus conference to revise the ASTRO definition of biochemical failure	Threshold value for PSA after radical radiotherapy	5
Pound et al.[21] (1999)	Retrospective review of a large surgical series (n = 1997) to examine the natural history of progression to distant metastases in men with elevated levels of PSA following surgery	Frequency of testing, threshold value for PSA after radical prostatectomy, natural history	5
Kuban et al.[18] (2006)	Primary study of radioisotopic implant as solitary treatment for localized prostate cancer (n = 2693); multiple definitions of PSA failure were tested for their ability to predict clinical failure	Threshold value for PSA after radical radiotherapy, natural history	4
ASTRO[16] (1997)	Consensus statement providing guidelines for PSA following radiation therapy	Frequency of testing, threshold value for PSA after radical radiotherapy	3
Cookson et al.[5] (2007)	AUA review of the variability in published definitions of biochemical recurrence; recommends a standard definition for patients whose cancer is treated with radical prostatectomy	Threshold value for PSA after radical prostatectomy, natural history	3
Horwitz et al.[19] (2005)	Determined the sensitivity and specificity of several definitions for biochemical failure using pooled data from 4839 patients whose cancer was treated with external-beam radiotherapy alone	Threshold value for PSA after radical radiotherapy	3
Stephenson et al.[20] (2006)	Tested 10 definitions of biochemical failure to identify the one that best explains metastatic progression; study involved 3125 patients who underwent radical prostatectomy	Threshold value for PSA after radical prostatectomy	3

Note: ASTRO = American Society for Therapeutic Radiology and Oncology, AUA = American Urological Association, PSA = prostate-specific antigen.

gen following radiotherapy from primary data;[34] the sensitivity and specificity of different definitions of biochemical failure were then compared, allowing for the control of characteristics that might affect the accuracy of definitions. More pertinently, statistical models using estimates of mean change and variability in a measurement over time to suggest optimal intervals for monitoring are being developed. A review of four case studies[35] found that for each topic (assessing cholesterol levels for secondary prevention of coronary artery disease,[36] long-term monitoring of blood pressure,[37] measurement of glycated hemoglobin in type 2 diabetes[38] and monitoring of CD4 cell counts in HIV-1 infection[39]) the results suggested frequent monitoring. There is clear potential for the extension of this work to monitoring in other settings.

Finally, general failings in the processes used to develop guidelines are likely to contribute substantially to the variations among published documents. In a review of guidelines for hypertension, Campbell and colleagues found a lack of rigour in the methods used to develop them.[29] In our sample, the National Institute for Health and Clinical Excellence and the Australian Cancer Network were the only organizations to cite published handbooks on the development of guidelines,[8,40] which may explain their higher ratings on the evaluation instrument we used. Guidelines that were clearly based on expert consensus tended to score considerably lower.[13,15] Savoie and colleagues[28] suggest that the greater the involvement of clinical experts in the development process, the less the recommendations reflect the evidence. It is likely that in the absence of clear methods for assessing monitoring strategies, greater involvement of methodologists on guideline panels would be beneficial.

Strengths and limitations

Our literature search was limited to a single, albeit large, medical database, supplemented with searches of more specialist resources, and records were limited to the English language. However, we have identified key guidelines that provide a good representation of the methods used by well-known agencies. Although other guidelines may be available, they are unlikely to have used alternative methods or to report on evidence that the included guidelines omitted.

Our use of the original Appraisal of Guidelines for Research and Evaluation instrument[39] may be criticized given that it has been 10 years since its publication; however, at the time the framework was chosen, the update to the original instrument[41] and other potentially useful frameworks[42,43] were not yet available. Nevertheless, our approach to assessing the development and content of the guidelines was systematic and provides a reasonable means of comparing guidelines.

We were not able to comment fully on the state of the underlying literature cited in support of monitoring schedules, as we did not retrieve copies of all primary studies cited. Furthermore, our "generous" approach to associating citations to recommendations may have inadvertently led to citations incorrectly being associated with recommendations. This may have led to some bias in favour of the guidelines, which could only be avoided by a full review of all evidence cited or by direct contact with the guideline's authors to determine which aspect of the recommendations were supported by the citations; neither of these solutions were within the scope of our review.

Our use of only one case study may limit the ability to generalize our results to other topics; however, we have no reason to believe the picture would be any better or worse in other areas. Indeed, Moschetti and colleagues found similar results for monitoring in cardiovascular disease.[26]

Our systematic approach to assessing the development and content of the guidelines provides insight into how strategies for monitoring are developed and reported, and we have presented a general picture of the type of evidence that has been cited. The true picture may be worse, given our attempt to attribute citations to recommendations wherever possible.

Conclusion

Our findings highlight the lack of a scientific or systematic approach to the development of monitoring schedules for prostate-specific antigen as reported in clinical guidelines, due both to inadequacies in the available evidence and its inappropriate use. This approach results in considerable inconsistency among guidelines.

Agencies producing guidelines should be encouraged to adopt systematic approaches to the development of their documents, such as those developed in the UK,[40] Australia[42] and the US,[43] and should take care to explicitly consider each element of a recommended monitoring schedule (interval, threshold and action to be taken on crossing the threshold) and the standard of its evidence base.

References

1. Glasziou PP, Aronson JK. An introduction to monitoring therapeutic interventions in clinical practice. In: Glasziou PP, Irwig L, Aronson JK, editors. *Evidence-based medical monitoring: from principles to practice*. Oxford (UK): Blackwell Publishing; 2008. p. 3-14.
2. Mant D. A framework for developing and evaluating a monitoring strategy. In: Glasziou PP, Irwig L, Aronson JK, editors. *Evidence-based medical monitoring: from principles to practice*. Oxford (UK): Blackwell Publishing; 2008. p. 15-30.
3. Irwig L, Glasziou PP. Choosing the best monitoring tests. In: Glasziou PP, Irwig L, Aronson JK, editors. *Evidence-based med-*

ical monitoring: from principles to practice. Oxford (UK): Blackwell Publishing; 2008. p. 63-74.
4. Lilja H, Ulmert D, Vickers AJ. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nat Rev Cancer* 2008;8:268-78.
5. Cookson MS, Aus G, Burnett AL, et al. Variation in the definition of biochemical recurrence in patients treated for localized prostate cancer: the American Urological Association Prostate Guidelines for Localized Prostate Cancer Update Panel report and recommendations for a standard in the reporting of surgical outcomes. *J Urol* 2007;177:540-5.
6. The AGREE Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. *Qual Saf Health Care* 2003;12:18-23.
7. The Royal College of Radiologists' Clinical Oncology Information Network, British Association of Urological Surgeons. Guidelines on the management of prostate cancer. *Clin Oncol (R Coll Radiol)* 1999;11:S53-88.
8. Australian Cancer Network Working Party on Management of Localised Prostate Cancer. *Clinical practice guidelines: evidence-based information and recommendations for the management of localised prostate cancer.* Canberra (Australia): National Health and Medical Research Council; 2002.
9. American Urological Association. *Guideline for the management of clinically localized prostate cancer*: *2007 update*. Linthicum (MD): American Urological Association Education and Research; 2007.
10. Dutch Urological Association. Prostate cancer. Nation-wide guideline version 1.0. Utrecht (Netherlands): Dutch Institute for Healthcare Improvement CBO; 2007.
11. National Cancer Institute. Prostate Cancer Treatment (PDQ®). Bethesda (MD): The Institute; 2008. Available: http://cancer.gov/cancertopics/pdq/treatment/prostate/HealthProfessional (accessed 2010 July 15).
12. National Institute for Health and Clinical Excellence. *Prostate cancer: diagnosis and treatment*. Cardiff (UK): National Collaborating Centre for Cancer; 2008.
13. American Urological Association. *Prostate-specific antigen best practice statement: 2009 update*. Linthicum (MD): American Urological Association Education and Research; 2009.
14. Heidenreich A, Bolla M, Joniau S, et al. EAU guidelines on prostate cancer. Arnhem (Netherlands): European Association of Urology; 2009.
15. National Comprehensive Cancer Network. NCCN clinical practice guidelines in oncology: prostate cancer. The Network; 2009.
16. American Society for Therapeutic Radiology and Oncology Consensus Panel. Consensus statement: guidelines for PSA following radiation therapy. *Int J Radiat Oncol Biol Phys* 1997;37:1035-41.
17. Roach M III, Hanks G, Thames H Jr, et al. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix Consensus Conference. *Int J Radiat Oncol Biol Phys* 2006;65:965-74.
18. Kuban DA, Levy LB, Potters L, et al. Comparison of biochemical failure definitions for permanent prostate brachytherapy. *Int J Radiat Oncol Biol Phys* 2006;65:1487-93.
19. Horwitz EM, Thames HD, Kuban DA, et al. Definitions of biochemical failure that best predict clinical failure in patients with prostate cancer treated with external beam radiation alone: a multi-institutional pooled analysis. *J Urol* 2005;173:797-802.
20. Stephenson AJ, Kattan MW, Eastham JA, et al. Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition. *J Clin Oncol* 2006;24:3973-8.
21. Pound CR, Partin AW, Eisenberger MA, et al. Natural history of progression after PSA elevation following radical prostatectomy. *JAMA* 1999;281:1591-7.
22. Sölétormos G, Semjonow A, Sibley PE, et al. Biological variation of total prostate-specific antigen: a survey of published estimates and consequences for clinical practice. *Clin Chem* 2005; 51:1342-51.
23. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189-202.
24. Vicini FA, Vargas C, Abner A, et al. Limitations in the use of serum prostate specific antigen levels to monitor patients after treatment for prostate cancer. *J Urol* 2005;173:1456-62.
25. Parker E, Glasziou P. Use of evidence in hypertension guidelines: should we measure in both arms? *Br J Gen Pract* 2009;59:e87-92.
26. Moschetti I, Brandt D, Perera R, et al. Accuracy of reporting monitoring regimens of risk factors for cardiovascular disease in clinical guidelines: systematic review. *BMJ* 2011;342:d1289.
27. Cruse H, Winiarek M, Marshburn J, et al. Quality and methods of developing practice guidelines. *BMC Health Serv Res* 2002;2:1.
28. Savoie I, Kazanjian A, Bassett K. Do clinical practice guidelines reflect research evidence? *J Health Serv Res Policy* 2000;5:76-82.
29. Campbell F, Dickinson HO, Cook JV, et al. Methods underpinning national clinical guidelines for hypertension: describing the evidence shortfall. *BMC Health Serv Res* 2006;6:47.
30. McAlister FA, van Diepen S, Padwal RS et al. How evidence-based are the recommendations in evidence-based guidelines? *PLoS Med* 2007;4:e250.
31. Burgers JS, Vailey JV, Klazinga NS, et al. Comparative analysis of recommendations and evidence in diabetes guidelines from 13 countries. *Diabetes Care* 2002;25:1933-9.
32. Glasziou PP, Irwig L, Aronson JK. *Evidence-based medical monitoring: from principles to practice*. 1st ed. Oxford (UK): Blackwell Publishing; 2008.
33. Bossuyt PM. Evaluating the effectiveness and costs of monitoring. In: Glasziou PP, Irwig L, Aronson JK, editors. *Evidence-based medical monitoring: from principles to practice.* Oxford (UK): Blackwell Publishing; 2008. p. 158-65.
34. Bellera CA, Hanley JA, Joseph L, et al. A statistical evaluation of rules for biochemical failure after radiotherapy in men treated for prostate cancer. *Int J Radiat Oncol Biol Phys* 2009;75:1357-63.
35. Stevens RJ, Oke J, Perera R. Statistical models for the control phase of clinical monitoring. *Stat Methods Med Res* 2010;19:394-414.
36. Glasziou PP, Irwig L, Heritier S, et al. Monitoring cholesterol levels: Measurement error or true change? *Ann Intern Med* 2008;148:656-61.
37. Keenan K, Hayen A, Neal BC, et al. Long term monitoring in patients receiving treatment to lower blood pressure: analysis of data from placebo controlled randomised controlled trial. *BMJ* 2009; 338:b1492.
38. Goldstein DE, Little RR, Lorenz RA, et al. Tests of glycemia in diabetes. *Diabetes Care* 2004;27:1761-73.
39. Buclin T, Telenti A, Perera R, et al. Development and validation of decision rules to guide frequency of monitoring CD4 cell count in HIV-1 infection before starting antiretroviral therapy. *PLoS ONE* 2011;6:e18578.
40. National Institute for Health and Clinical Excellence. *The guidelines manual*. London (UK): NICE; 2009.
41. Brouwers M, Kho ME, Browman GP et al. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ* 2010 ;182:E839-42.
42. Hillier S, Grimmer-Somers K, Merlin T, et al. FORM: an Australian method for formulating and grading recommendations in evidence-based clinical guidelines. *BMC Med Res Methodol* 2011;11:23.
43. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. *Clinical practice guidelines we can trust*. Washington (DC): The National Academies Press; 2011.

Affiliations: From the Department of Public Health, Epidemiology and Biostatistics (Dinnes, Deeks) University of Birmingham, Edgbaston, Birmingham, UK; Leeds Institute for Health Sciences (Hewison), University of Leeds, Leeds, UK; Centre for Statistics in Medicine (Altman), University of Oxford, Oxford, UK

Contributors: Jonathan Deeks, Douglas Altman and Jenny Hewison conceived the idea for the article. Jacqueline Dinnes performed the literature search, analyzed and interpreted the data and drafted the article. All of authors critically revised the manuscript for important intellectual content and gave their final approval of the version submitted for publication. Jonathan Deeks is the guarantor.

Funding: This publication presents independent research commissioned by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10101). The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health. For further details, please refer to the website www.biomarkerpipeline.org. All of the researchers involved in this project are independent of the funding body.

Acknowledgement: The authors thank Dr. Rik Bryan, University of Birmingham, for helpful comments.