Variation in subfertility care measured by guideline-based performance indicators

S.M. Mourad1,2, W.L.D.M. Nelen1, R.P.M.G. Hermens2, L.F. Bancsi3, D.D.M. Braat1, G.A. Zielhuis4, R.P.T.M. Grol2 and J.A.M. Kremer1,5

1Department of Obstetrics and Gynaecology, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands; 2Centre for Quality of Care Research (WOK), Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB, Nijmegen, The Netherlands; 3Department of Obstetrics and Gynaecology, Rijnstate Hospital, PO Box 9555, 6800 TA Arnhem, The Netherlands; 4Department of Epidemiology and Biostatistics, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands
5Correspondence address. E-mail: j.kremer@obgyn.umcn.nl

BACKGROUND: About 30–40% of patients do not receive care based on available scientific evidence. For subfertility, this may imply unnecessary and expensive diagnostic tests and treatments. It is therefore important to identify gaps in performance by monitoring current subfertility care. A set of 39 guideline-based performance indicators was previously developed for this purpose. This study aimed to assess several quality criteria of the indicator-set and to use the set to assess current subfertility care. METHODS: A historic cohort study was performed in 16 Dutch subfertility clinics; 2698 couples were invited to participate. Indicator data were gathered by medical record extraction, and patient and professional questionnaires. Quality criteria for each indicator (measurability, reliability, applicability, improvement potential, discriminatory capacity, complexity and case-mix stability) were assessed. Current practice was measured as adherence to the separate indicators. RESULTS: One thousand four-hundred and ninety-nine (56%) couples participated. All indicators were measurable, but the results for the other quality criteria varied. In total, 14 of the 39 indicators scored <50% adherence. Variation in performance between the clinics was up to 100%. The highest median adherence (86%) is found within the guideline ‘indications for IVF-treatment’. The lowest median adherence is found within the guideline ‘anovulation’ (44%). CONCLUSIONS: This study shows the quality of the developed indicator-set for monitoring clinical subfertility care. A first assessment in the Netherlands reveals large variation between clinics and ample room for improvement of care.

Keywords: guidelines; performance indicator; monitoring; quality of care; subfertility

Introduction
According to medical literature, ~30–40% of patients do not receive care based on available scientific evidence (McGlynn et al., 2003). On top of this, an estimated 20–25% of provided health care is unnecessary (McGlynn, 1995; Schuster et al., 1998; Grol, 2001). In the case of subfertility, this could mean the use of unnecessary and expensive diagnostic tests and assisted reproduction technologies (ART) or the realization of complicated high-order pregnancies. Subsequently, this may have substantial physical and psychological consequences for the patients involved (Greil, 1997; Kallen et al., 2005; Ombelet et al., 2006; Cousineau and Domar, 2007; Schieve et al., 2007). Patients should receive the best achievable care with a minimum of complications and at minimal costs. To improve both outcome and process of health care, clinical practice guidelines have been developed summarizing the best available evidence (Grimshaw et al., 2004). However, such guidelines do not implement themselves (Grol, 1997). Efforts to monitor actual health care and to identify gaps in performance show where improvement is needed. Guideline-based performance indicators are useful tools for monitoring health care (Grol et al., 2002). Consequently, they can be used to assess changes in clinical practice as well, which makes them indispensable instruments within clinical improvement programmes.

We reported previously on the development of a guideline-based set of 39 process and structure indicators for the entire spectrum of clinical subfertility care by means of a systematic six-step RAND-modified Delphi method (Mourad et al., 2007). The selected indicators focus on process and structure of subfertility care rather than outcome. Although current international debate focuses mainly on outcome measures (Dickey...
et al., 2004; Schieve and Reynolds, 2004), process indicators are valuable instruments within performance assessment and improvement programmes because they reveal exactly where changes in care might be necessary (Palmer, 1997; Eddy, 1998; Brook et al., 2000; Mant, 2001; Rubin et al., 2001b; Mourad et al., 2007; Nelen et al., 2007). This does not mean, however, that clinical outcomes are unimportant, but to change outcomes, it is necessary to first initiate changes in the process and structure of care. The primary aim of this study was to test the indicator-set for several quality criteria (Terwee et al., 2007). Such a practice test is necessary to demonstrate its value as an instrument for monitoring and improvement of clinical performance (Rubin et al., 2001a; Campbell et al., 2003; Schouten et al., 2005; Hermens et al., 2006; Ouwens et al., 2006). The second aim, conditional on sufficient quality of the instrument, was to assess the variation in current subfertility care in a large sample of clinics.

Materials and Methods

Study design and study population

We conducted a retrospective cohort study in 16 Dutch clinics using medical record and questionnaire data. The study was approved for all clinics by ‘the Regional Review Board for Human Research (CMO) Arnhem-Nijmegen (CMO no. 2004/193’). To assess current clinical subfertility care in the Netherlands, we aimed at including a broad patient cohort from the participating clinics. The clinics’ characteristics varied to ensure that delivered care was representative for Dutch standards. There was one academic and one tertiary care clinic and seven clinics offered secondary care (IVF/ICSI-treatment). These clinics were all also teaching clinics and of large or intermediate size; the other seven clinics were smaller, non-teaching facilities. In total, 15 clinics were national health, and one of the smaller secondary care clinics was a private clinic.

To include a representative patient group, potential participating couples were randomly selected by means of each clinics’ financial DBC (Diagnosis/Treatment Combination code) registration database. In this national registration, patients undergoing diagnostics or treatment for subfertility are identified with a specific Fertility-code (F-code). Couples were apt for inclusion if they had an active F-code anytime between 1 April and 30 June 2005. The assessment focused on the care they received in the period 1 January–30 June 2005. In each clinic, a random sample of eligible couples (50–500 depending on the size of the clinic) was invited to participate in the study. They were sent an informed consent form and a questionnaire: the former included consent to use their medical data from the clinic: Couples who had insufficient knowledge of the Dutch language or did not visit the subfertility clinic in the study period were excluded. In total, 2698 couples were invited.

The clinical performance indicators

We used a set of 39 clinical performance indicators for subfertility care, which was developed using a rigorous and systematic six-step approach. The set was based on literature, existing international guidelines and indicators and 10 Dutch (evidence and consensus based) subfertility guidelines (Mourad et al., 2007). The latter includes the Dutch Embryo Act and nine subfertility guidelines of the Dutch Society of Obstetrics and Gynecology (NVOG): initial assessment of fertility, anovulation and child wish, male subfertility, tubal pathology, endometriosis, premature ovarian failure (POF), intrauterine insemination (IUI), indications for IVF treatment and ovarian hyperstimulation syndrome (OHSS).

The indicator-set encompasses 37 process indicators and 2 structural indicators, and comprises the entire spectrum of subfertility care: the initial assessment of fertility, diagnostics, treatments and possible complications. Indicators were operationalized by the construction of a numerator and denominator, each consisting of several variables. The numerator is formed by the proportion of patients in which there is adherence to a recommendation and the denominator by the proportion of patients in which the recommendation is applicable. The indicators within the guideline ‘initial assessment of fertility’ are shown in Table I. The remaining indicators, ranked per guideline, are shown in Supplementary Table Ia–e in the web-based version of this article.

Data collection to assess quality criteria and current practice

Data for indicators regarding indications for ART (n = 7), diagnostic procedures (n = 6) and treatment procedures (n = 20) were extracted from the individual medical records by three trained investigators. Data for indicators regarding information provision to patients (n = 4) were collected by means of a patient questionnaire and data for indicators regarding the structure of care (n = 2) by a short professional questionnaire. The patient questionnaire was piloted in a group of 30 subfertile couples and this led to minor adjustments in formulation of the questions before it was used in the total study population.

Definitions and analysis

The application of a systematic and rigorous consensus method for indicator development results in high face and content validity of this indicator-set. We measured the following quality criteria of all indicators in the set, in order to demonstrate its value as an instrument for assessment of subfertility care (Streiner and Norman, 1994; Rubin et al., 2001a; Grol et al., 2002; Campbell et al., 2003; Grol et al., 2005; Schouten et al., 2005; Ouwens et al., 2006; Wollersheim et al., 2007).

Interested readers are referred to Supplementary Table II in the web-based version of this article which contains a list of suggested additional reading on the use and validity of (process) indicators in clinical quality and performance assessment.

(i) Measurability: an indicator is considered measurable if data to fill the numerator and denominator of the indicator can be made available through data collection (e.g. medical records, complication or treatment databases or a survey) (Rubin et al., 2001a).

(ii) Reliability: if the measurement of indicator data by two different data-collectors is reliably comparable, there is high reliability (Rubin et al., 2001a). Two independent reviewers abstracted a random sample of 10% (n = 32) of medical records from two participating clinics. The extent of agreement between these data reviewers on the level of process indicator scores, corrected for chance, was calculated using Cohen’s kappa coefficient of inter-rater reliability, Coefficients of 0.4–0.6 represent moderate inter-rater reliability, coefficients of 0.61 or higher are considered very good (Landis and Koch, 1977).

(iii) Applicability: for accurate performance assessment, an indicator is preferably applicable to a substantial proportion of reviewed patients (>10 cases); this is referred to as applicability (Grol et al., 2002).

(iv) Improvement potential: when an indicator is used to detect changes in clinical performance, it is a prerequisite that improvement is possible at all; if overall performance for a certain indicator is already very high, the indicator has no improvement potential. We defined improvement potential as an overall performance score of <90% (Ouwens et al., 2006).

(v) Discriminatory capacity: the discriminatory capacity indicates whether an indicator is able to discriminate practice
Table I. Adherence to indicators for clinical subfertility care: an assessment in 16 Dutch fertility clinics.

<table>
<thead>
<tr>
<th>Initial assessment of fertility</th>
<th>Range and median adherence for 16 clinics (%)</th>
<th>Median adherence (43)</th>
<th>Range</th>
<th>n patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>The initial fertility assessment should result in both a diagnosis and a prognosis</td>
<td>23</td>
<td>0–100</td>
<td>428</td>
<td></td>
</tr>
<tr>
<td>The initial fertility assessment should consist of three parts: semen analysis, tubal patency assessment and cycle analysis</td>
<td>31</td>
<td>0–83</td>
<td>179</td>
<td></td>
</tr>
<tr>
<td>Couple’s history-taking should cover at least:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Age of both partners</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Duration of subfertility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Type of couple’s subfertility (primary or secondary)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman’s physical examination should include assessment of the body mass index</td>
<td>98</td>
<td>83–100</td>
<td>361</td>
<td></td>
</tr>
<tr>
<td>Life-style advice concerning bodyweight, smoking, and the alcohol and drug use should be part of the counselling regarding pregnancy-probabilities</td>
<td>100</td>
<td>100</td>
<td>249</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Median adherence of all indicators within one guideline.*
performance between different hospitals. High discriminatory capacity is therefore present when the range in scores between the lowest- and highest-scoring clinics is $>20\%$. (Grol et al., 2002).

(vi) Complexity: the concept of complexity was used for the number of variables needed to fill the numerator and denominator of an indicator. Complexity is thus a measure that reflects the amount of investments needed to assess an indicator; the higher the complexity the more effort it takes to measure an indicator. The maximum value for complexity was beforehand set at five different variables; in the case of complexity scores of more than five variables, subdivision of the indicator was considered. Variables relating to the inclusion of patients were not counted because those variables can vary per clinic and depend on design of the medical records. For the indicators measured by medical record extraction, two researchers (W.L.D.M.N. and S.M.M.) counted independently the number of variables to fill the numerator and denominator of each indicator. For the indicators measured by means of a professional or patient questionnaire, we counted the number of answers needed to determine the indicator. If there was disagreement between the two researchers’ findings, discussion took place to reach an agreement.

(vii) Case-mix stability: case-mix stability is an important asset that enables the application of an indicator for monitoring within a specific hospital over time or to compare hospitals of different sizes and settings (Rubin et al., 2001a). Therefore, the relationship between certain patient-characteristics that may vary considerably between clinics [‘woman’s age’, ‘type of subfertility (primary or secondary)’ and ‘duration of subfertility’] and the indicator scores was analysed to decide whether correction for case mix is necessary.

Analysis

Collected data were entered in a database using the Statistical Package for the Social Sciences (SPSS 14.0 for Windows®, SPSS Inc., Chicago, IL, USA). Descriptive analyses were performed for each indicator, and current practice was expressed as percentage of adherence to an indicator. The median adherence of each separate clinic was calculated, as well as the adherence of all clinics together.

Results

Response

The mean participation rate was 56% (1499 of 2698 invited couples), varying between 47% and 72% per clinic. Demographic characteristics of the participating couples are shown in Table II. Mean female age was 33 years and male age 35 years. Forty-seven couples (3.2%) were of non-Dutch origin. Of the participating couples, 73% suffered from primary subfertility. The median duration of subfertility was 38 months. A total of 53% had a high education level and 72% had a more than modal income (gross $>€2750/month).

Assessment of quality criteria of the indicator-set

Table III shows the quality criteria of the indicators, within the guideline ‘initial assessment of fertility’. Supplementary Table IIIa–i, available in the web-based version of this article, shows these criteria for the remaining guidelines. All indicators were measurable and made available through

| Table II. Baseline characteristics of the participating couples ($n=1499$). |
|-------------------------------|-------------------|
| Characteristics               | % of couples      |
| Mean age in years (SD)         |                   |
| Female                        | 32.8 years (4.1)  |
| Male                          | 35.1 years (5.0)  |
| Ethnic background$^a$          |                   |
| Dutch                         | 96.8              |
| Non-Dutch                     | 3.2               |
| Gross monthly family income (euros)$^b$ |
| $<1100$                       | 1.6               |
| 1100–1760                     | 4.6               |
| 1760–2750                     | 22.2              |
| $>2750$                       | 71.6              |
| Education level per couple$^c$ |
| Low                           | 6.4               |
| Intermediate                  | 40.7              |
| High                          | 52.9              |
| Type of subfertility$^d$       |                   |
| Primary                       | 72.8              |
| Secondary                     | 27.2              |
| Median duration of subfertility in months (SD)$^e$ |
| Low (18–25)                   | 14.0              |
| Normal (18–25)                | 53.9              |
| High (>25)                    | 32.0              |

$^a$Ethnic background of the couples was determined by the origin of both partners: Dutch, one or both partners are of Dutch origin; non-Dutch, both partners are not of Dutch origin.

$^b$Gross monthly family income was categorized according to social security standards and modal income in euros: $<1100$ less than Dutch modal income; 1100–1760 Dutch modal income; 1760–2750 up to 1.5 $×$ Dutch modal income; $>2750$ more than Dutch modal income.

$^c$Education level of the couples was determined by the highest education level of both partners: low, primary or lower vocational education; intermediate, secondary or intermediate vocational education; high, higher professional education or university.

$^d$Type of subfertility was determined for the couple.

$^e$Duration of subfertility was defined as the period between the start of regular unprotected sexual intercourse and 1 January 2005, the beginning of the study period.

either medical record extraction, or a patient or professional questionnaire. The reliability is high for questionnaire data, because it is first-hand information. For medical record data, reliability was substantial, reflected in an average Cohen’s kappa coefficients of 0.86 (range 0.48–1.0). In total, 18 indicators were inapplicable (<10 patients). For these indicators, the other quality criteria could not be assessed. A total of eight indicators had adherence scores of 90% or more, and therefore possess too little room for improvement. Regarding discriminatory capacity, a total of 11 indicators scored $<20\%$. This understandably includes the eight indicators that had no room for improvement. None of the indicators required more than one data source. Complexity values ranged mostly from 1 to 5. An exception was the indicator concerning ‘the items that should be discussed during an intake prior to IVF/ICSI-treatment’. This indicator encompassed many topics and concerned 18 variables and was therefore subdivided into five different parts for accurate analysis. Assessment of case-mix stability revealed that the following three indicators are in need of correction for ‘duration of subfertility’: (i) ‘the initial fertility assessment should consist of three parts: semen-analysis, tubal patency assessment and cycle analysis’, (ii) ‘life-style advice concerning bodyweight, smoking and
the alcohol and drug use should be part of the counselling regarding pregnancy-probabilities’ and (iii) ‘in the case of unexplained subfertility in a woman <36 years, there is an indication for IVF after 3 years of subfertility’ needs correction for ‘duration of subfertility’, ‘woman’s age’ and ‘type of subfertility’ when the analysis takes place. All other indicators showed equal distribution of these patient-characteristics.

The overall results of this quality criteria assessment are summarized in Table IV.

Table IV. Overall scores of the indicators (n = 39) for each quality criterion.

<table>
<thead>
<tr>
<th>Quality criteria</th>
<th>Number of indicators that met the criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurability</td>
<td>39</td>
</tr>
<tr>
<td>Reliability</td>
<td>39</td>
</tr>
<tr>
<td>Applicability</td>
<td>21</td>
</tr>
<tr>
<td>Improvement potential</td>
<td>31</td>
</tr>
<tr>
<td>Discriminatory capacity</td>
<td>28</td>
</tr>
<tr>
<td>Complexity</td>
<td>38</td>
</tr>
<tr>
<td>Case-mix stability</td>
<td>35</td>
</tr>
</tbody>
</table>

Current practice: adherence to indicators

Table I shows the median clinical performance and the range of performance of the clinics for each indicator within the guideline ‘initial assessment of fertility’. Supplementary Table Ia–e contains the results for the remaining indicators and is available in the web-based version of this article. The indicators regarding endometriosis, OHSS, tubal pathology and POF (Mourad et al., 2007) were applicable to too few patients (n < 10) in this sample and are therefore not shown in Table I. From Table I and Supplementary Table Ia–e, it can be concluded that the median adherence of 14 indicators was <50%. Of those, six indicators even scored <25%. The median adherence to the separate indicators varied from 0% (e.g. ‘in the case of abnormal semen-analysis, a complete andrological history-taking, physical examination and one extra semen-analyses should be performed’) up to 100% (e.g. ‘woman’s physical examination should include an assessment of the body mass index’). The variation between clinics differed widely and is reflected in the range of adherence within one indicator. We found, for example, on the one hand exactly equal adherence between all clinics (e.g. all clinics scored 100%) for the indicator ‘routine use of hCG for luteal support after IVF is not recommended’ and on the other hand a 100% difference between the clinics (e.g. two clinics scoring 0% and 100%, respectively) for the indicator ‘initial fertility assessment should result in a prognosis and a diagnosis’.

Finally, the results show that the highest median adherence (86%) is found within the guideline ‘indications for IVF-treatment’. The lowest mean adherence is found within the guideline ‘initial assessment of fertility’ (43%), followed closely by the guideline ‘anovulation’ (44%).

Discussion

This study demonstrates several quality criteria of a set of 39 guideline-based process and structure indicators for comprehensive clinical subfertility care. Moreover, a first assessment of current subfertility care in the Netherlands discloses a large variation in the performance between clinics.

During the original indicator-selection procedure, the expert panels were asked to take five criteria into account when selecting key recommendations from the guidelines (Mourad et al., 2007). In this practice test, we assessed additional quality criteria of the proposed indicators. We can conclude that most of them meet these criteria, with several remarkable results. Improvement potential and applicability were not always high. Apparently, a considerable discrepancy exists between the experts’ estimate of these criteria on the one hand and currently provided care on the other hand. This underlines that even a rigorous indicator-development procedure should be complemented with a practice test. On the basis of the results of our practice test, particularly applicability and improvement
potential (Tables III and IV and Supplementary Table IIIa–i), we can distinguish three subsets within our original set of 39 performance indicators. Indicators within the ‘first subset’ score highly for applicability and improvement potential and can therefore be used for quality monitoring purposes and as a baseline measurement for improvement programmes in the involved clinics. A ‘second subset’ contains indicators with low improvement potential, which makes them useless for improvement programmes, but with high applicability, which makes them still suitable for monitoring purposes, namely to ensure that adherence continues to be high in the future. Within the ‘third subset’, we find indicators with low applicability. They concern relatively rare conditions or complications (e.g. POF and severe OHSS), so inclusion numbers within our random patient sample were too small for adequate indicator assessment. This does not mean, however, that this subset needs to be discarded immediately for any future use; but the inclusion of enough cases for such indicators requires a more specific patient sample, e.g. drawn from separate complication registrations or large (national) databases.

The question arises whether this set of indicators is also suitable for countries other than the Netherlands. Only 2 out of the 39 indicators (e.g. laboratory accreditation by a Dutch committee and the maximum number of two embryos transferred per cycle) may be more or less specific for Dutch practice, but both can be easily adapted for international use by judging not the content but the topic of the indicator. Generalizability is thus not a barrier for international adaptation of the presented indicator-set.

Adherence to the guideline recommendations

Our study shows that adherence to guideline-based indicators ranged widely per clinic, per guideline and per individual indicator. Adherence to the indicators is relatively poor; 14 indicators score below 50% adherence. This means there is ample room for improvement of clinical subfertility care. Professionals should be urged to improve the implementation of those guideline-recommendations that showed poor adherence in their clinic. A 100% adherence score may not always be a viable goal, because there can in specific cases be good reasons to divert from guideline-recommendations. However, professionals should at least aim to rival the best scoring clinic, which is a realistic benchmark. The lowest adherence (less than 50%) is found within the guidelines ‘initial assessment of fertility’ and ‘anovulation’; priority should therefore be given to implementation of these guidelines. It is furthermore important to realize that a low score on an indicator does not automatically mean that there is a problem in the quality of care; it is, however, a signal to further estimate the matter in order to try to understand underlying causes and processes.

There are some results that stand out in particular. Several of the critical indicators with low adherence relate to communication and patient information. Again, the broad variation in adherence to these indicators and the fact that some clinics do score highly, illustrate that these results cannot be simply disregarded or attributed to recall bias (Reading, 1981; Kessels, 2003). These findings are in line with previously conducted qualitative and quantitative research, identifying physicians’ ‘lack of self-efficacy regarding communication’ and ‘low outcome expectancy’ as main barriers for adherence to a subfertility guideline (Haagen et al., 2005). However, this is plainly disturbing from the perspective of patient-centred care, as former research clearly discloses subfertile patients’ preferences for comprehensible and complete information (Schmidt, 1998; Souter et al., 1998; Malin et al., 2001; Schmidt et al., 2003). Increased awareness among care-providers of these patient preferences and feedback on low indicator scores should lead to an improvement in information provision and transparency of care. Another low-scoring indicator is the one recommending unstimulated over stimulated IUI, in the case of unexplained infertility, to prevent multiple pregnancies (15% adherence, range between clinics 0–48%). This perfectly reflects the situation where consensus on a specific topic is lacking. None of the clinics score 100% adherence, meaning this recommendation is far from consistently followed. This might be because both patients and professionals are apt to change their treatment policy and are willing to take more risks, e.g. after previous unsuccessful cycles of unstimulated IUI. This probably also explains the poor adherence to the recommendation ‘mono-ovulation should be the result of ovulation-induction’. However, it is the responsibility and moral duty of the professional to guard patient safety during any treatment, and 0% scores should therefore be taken very seriously. Fortunately, more reticence is seen in the high adherence (79%) of the recommendation ‘IVF-treatment in women <36 years is only indicated after 3 years of subfertility’. This reflects the professionals’ commitment to prevent over-treatment and strive for cost-effectiveness and efficacy of this expensive and intensive type of subfertility treatment, by respecting the possibility of less invasive approaches and treatment-independent pregnancies (Goverde et al., 2000; Evers, 2002).

This study also has some limitations. Indicator measurement turned out to be quite a laborious exercise, including widely distributed questionnaires and an extensive medical record search to complete indicator data. Poor availability of medical data and the lack of adequate data resources are common problems in performance assessment efforts (Gagliardi et al., 2005; Pronovost et al., 2007), but this should never be the sole reason to discard rigorously selected performance indicators. On the contrary, care-providers and policy-makers should undertake efforts to overcome these barriers. The ongoing introduction of electronic patient records offers great opportunities for collaboration with clinical improvement programmes. It is therefore important to conduct further research to uncover which relatively small investments or adjustments to existing databases are necessary to facilitate monitoring of current care.

In conclusion, we demonstrated the quality of a developed set of guideline-based indicators for comprehensive clinical subfertility care. A practice test in the Netherlands revealed large variation and ample room for improvement of subfertility care. Such objective assessment of care can help professionals to identify and subsequently target the domains of care in need of improvement. The next step will be the translation of indicator results into appropriate implementation strategies that...
aim to bridge the gaps between the best available evidence we present in our guidelines and the care we deliver in daily practice.

Supplementary Data
Supplementary data are available at http://humrep.oxfordjournals.org/

Acknowledgements
We like to acknowledge all patient couples for their participation and the 16 subfertility clinics for their cooperation in data sampling. Special thanks to the coordinating gynaecologists of these clinics: P.A. van Dop (St Catharina Ziekenhuis, Eindhoven), B.W. Mol (Maxima Medisch Centrum, Veldhoven), A.M.H.W. Franssen (Canius Wilhelmina Ziekenhuis, Nijmegen), M. Van Rozendaal (Ziekenhuis Bernhoven, Veghel), J.H.A. Vollebergh (Ziekenhuis Bernhoven, Oss), A.G. Minkhorst (Maasziekenhuis Pantein, Boxmeer), C.J.C.M. Hamilton (Jeroen Bosch Ziekenhuis, den Bosch), H. Ruis (behandelcentrum stichting Geertgen, de Mortel), G.J.J.M. Muijsers (Rivierenland Ziekenhuis, Tiel), J.H. Schagen van Leeuwen (St Antonius Ziekenhuis, Nieuwveen), L.J. van Dam (Gelre Ziekenhuis, Apeldoorn), E. Scheenjes (Ziekenhuis Gelderse Vallet, Ede), A.P.E. Schmoutziguer and L.F.Bancsi (Ziekenhuis Rijnstate, Arnhem), V.M. Blom (Streekziekenhuis Zevenaar, Zevenaar), E.M. Tepe, C.J.C.M. sius Wilhelmina Ziekenhuis, Nijmegen), M. Van Rozendaal (Ziekenhuis, Nieuwegein), L.J. van Dam (Gelre Ziekenhuis, Apeldoorn), E. Scheenjes (Ziekenhuis Gelderse Vallet, Ede), A.P.E. Schmoutziguer and L.F.Bancsi (Ziekenhuis Rijnstate, Arnhem), V.M. Blom (Streekziekenhuis Zevenaar, Zevenaar), E.M. Tepe. (Slingeland Ziekenhuis, Doetinchem). We also thank, medical student Evelyn Verheijden and research assistant Janine Liefers for their help in data management and analysis.

Funding
The Netherlands Organisation for Health Research and Development (ZonMw) funded this project (Grant No. 945-14-116).

References
Dickey RP, Sartor BM, Pyrzak R. What is the most relevant standard of success in assisted reproduction?: no single outcome measure is satisfactory when evaluating success in assisted reproduction; both twin births and singleton births should be counted as successes. Hum Reprod 2004;19:783–787.
Schieve LA, Reynolds MA. What is the most relevant standard of success in assisted reproduction?: challenges in measuring and reporting success rates for assisted reproductive technology treatments: what is optimal? Hum Reprod 2004;19:778–782.


Submitted on March 25, 2008; resubmitted on June 19, 2008; accepted on June 25, 2008