Technical and performance characteristics of anti-Müllerian hormone and antral follicle count as biomarkers of ovarian response

Stamatina Iliodromiti1, Richard A. Anderson2, and Scott M. Nelson1,*

1School of Medicine, University of Glasgow, Glasgow G31 2ER, UK 2MRC Centre for Reproductive Health, University of Edinburgh, Edinburgh EH16 4TJ, UK

*Correspondence address. School of Medicine, University of Glasgow, Room 2.52 Level 2, New Lister Building, Glasgow Royal Infirmary, Glasgow G31 2ER, UK. Tel: +44-141-201-8581; E-mail: scott.nelson@glasgow.ac.uk

Submitted on July 15, 2014; resubmitted on October 24, 2014; accepted on November 7, 2014

TABLE OF CONTENTS

- Background
  - The physiology of follicle growth determining availability for exogenous gonadotrophin recruitment
- Methods
  - Anti-müllerian hormone
    - Factors that can influence AMH values
  - Antral follicle count
    - Factors that can influence AFC
- Comparison of AMH and AFC performance in predicting ovarian response
- Conclusion

BACKGROUND: Stratified (individualized) medicine has been recognized as a key priority for policy makers and healthcare providers. The main principle of individualized care depends on utilizing patients’ characteristics and biomarkers to predict prognosis, tailor intended treatment and predict treatment outcomes. In reproductive medicine a wide variety of biomarkers have been proposed as predictors of ovarian response; of these, anti-Müllerian hormone (AMH) and antral follicle count (AFC) are purported as exhibiting the most favourable analytical and performance characteristics. Previously AFC and AMH have been considered essentially interchangeable; however, recent trial data have questioned this postulation. The aim of this review is to present an analysis of the strengths and weaknesses of these biomarkers as predictors of ovarian response, using both physiological and technical perspectives.

METHODS: We have conducted a systematic search of the most recent (to May 2014) relevant literature and summarized the existing evidence. Articles written in a language other than English without an available English translation were excluded.

RESULTS: Both AMH values and AFC can be influenced by comparable technical, physiological and exogenous factors. AMH displays some variation within and between cycles, consistent with its physiological role in follicle development, and there are growing data on the impact of pharmacological treatments and pathological conditions but cycle-independent measurement is appropriate for clinical purposes. A range of issues with manual AMH assays may be resolving with the development of fully automated assays. Despite described standardization of its measurement technique, AFC is subject to marked inter- and intra-operator variability and the effects of external influences are likely to be comparable. Outwith some highly specialist centres, the intracyclical variation in AFC requires its measurement between Day 2 and 4 of the cycle. Observational studies suggest comparable performance characteristics for AMH and AFC in predicting poor and high ovarian response, but recent RCTs suggest markedly better performance for AMH.

CONCLUSIONS: The performance characteristics of both AMH and AFC for the prediction of ovarian response to exogenous gonadotrophins have been inflated by single site observational cohorts, resulting in the viewpoint that AMH and AFC exhibit equivalent performance.
characteristics. Large scale multicentre RCTs, with centralized assay performance, have demonstrated that AMH is substantially the more accurate and robust biomarker, probably reflecting difficulties with standardization of AFC determination. While AFC retains some advantages, particularly immediacy and accessibility, international standardization of AMH combined with a stable automated assay is likely to enhance its performance as the biomarker of choice in predicting the ovarian response in assisted conception.

**Key words:** anti-Müllerian hormone / antral follicle count / inter-individual variation / intra-individual variation / ovarian response

### Background

Stratified medicine is recognized as a key global priority for healthcare providers, patients, and pharmaceutical and diagnostic industries. Achieving personalized care with provision of the ‘right treatment, for the right person, at the right time’ should be an inevitable progression as we gain greater understanding of the aetiology and pathophysiology of disease but requires critical assessment of all aspects of care. Advances in understanding have enabled us to predict disease reliably at population levels, with existing and novel biomarkers now being evaluated for incorporation into composite models (Tunstall-Pedoe, 2011). Reproductive medicine has taken a notable lead in the use of prognostic models for the right person, at the right time’ should be an inevitable progression with differential rates of follicular activation at different ages is necessary to have a continuous supply of growing follicles to support the selection processes that precede ovulation (Wallace and Kelsey 2010) and is probably influenced by health status. Secondly follicles undergo atresia at all stages of development (Zuccotti et al., 2011). Thirdly the number of activated follicles reflects the total pool of primordial follicles in a variable manner, with markedly different correlation coefficients in childhood and adult life (Kelsey et al., 2012). Lastly ovarian reserve depletion will depend on the initial quantity of primordial follicles, and the rate of primordial follicle recruitment. Collectively this means that although in adult life, biomarkers of activated follicles such as AMH and AFC can potentially reflect the primordial follicle pool (Hansen et al., 2011), their greatest strength and value will be in indicating the number of follicles that are at late stages of follicular development and capable of responding to exogenous gonadotrophins. Thus AMH and AFC are of greatest value in reflecting what has been termed the functional rather than the true ovarian reserve (Anderson et al., 2012).

### Methods

Multiple strategies were used to identify relevant demographic, epidemiological, clinical and biological studies relevant to the broad topic of AFC and AMH, without a date limit and up to May 2014. We searched in sociological online libraries (IBSS, SocINDEX), MEDLINE, EMBASE, EMBASE CLASSIC, Clinicaltrials.gov, ISRCTN registry, EU Clinical Trials Register, UMIN-CTR, ANZCTR, PubMed and Google Scholar using the following key words and hierarchical MeSH terms; anti-müllerian hormone, AMH, müllerian inhibiting factor, MIF, antral follicle count, AFC. Additional journal articles were identified from the bibliography of studies included as well as textbooks and hand searches of other source materials including conference proceedings. Articles written in a language other than English without an available English translation were excluded from our review. From this, we identified and focused on key topics (listed in Table of Contents) where it was judged that there had been clinically relevant advances in the understanding of ovarian response prediction with implications for improved diagnostics and prediction models.

### Anti-müllerian hormone

#### Factors that can influence AMH values

**Assay technical issues**

Since the original reports of measurement of serum AMH in 1990 (Baker et al., 1990; Hudson et al., 1990; Josso et al., 1990), there has been continual development of the immunoassay by a variety of companies, utilizing different antibody pairs. At present four manual enzyme-linked immunosorbent assays are available and the performance characteristics of these assays are summarized in Table I. Although these assays exhibit good within laboratory reproducibility, they display substantial variability between laboratories, due to the lack of automation and site-specific
Table 1  Anti-Müllerian hormone (AMH) assays characteristics.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Imprecision</td>
<td>&lt;8%</td>
<td>&lt;14%</td>
<td>&lt;6%</td>
<td>&lt;6%</td>
<td>1.8—2.0%</td>
<td>2.87—4.34%</td>
</tr>
<tr>
<td>Sample type</td>
<td>Serum, plasma</td>
<td>Serum, plasma</td>
<td>Serum, plasma</td>
<td>Serum, plasma</td>
<td>Serum, Li-heparin plasma</td>
<td>Serum, Li-heparin plasma</td>
</tr>
<tr>
<td>Minimum sample volume</td>
<td>20 µl</td>
<td>25 µl</td>
<td>50 µl</td>
<td>100 µl</td>
<td>50 µl</td>
<td>20 µl</td>
</tr>
<tr>
<td>Incubation time</td>
<td>&lt;3 h</td>
<td>3 h</td>
<td>2.5 h</td>
<td>4.5 h</td>
<td>18 min</td>
<td>39 min</td>
</tr>
<tr>
<td>Limit of detection (LoD)</td>
<td>0.08 ng/ml</td>
<td>0.14 ng/ml</td>
<td>0.023 ng/ml</td>
<td>0.0012 ng/ml</td>
<td>0.01 ng/ml</td>
<td>≤0.02 ng/ml</td>
</tr>
<tr>
<td>Limit of Quantification (LoQ)</td>
<td>0.16 ng/ml</td>
<td>0.35 ng/ml (Decanter et al., 2014)</td>
<td>0.06 ng/ml</td>
<td>0.0039 ng/ml</td>
<td>0.03 ng/ml</td>
<td>≤0.08 ng/ml</td>
</tr>
<tr>
<td>Measurement range</td>
<td>0.16–22.5 ng/ml</td>
<td>0.42–21.0 ng/ml</td>
<td>0.06–11.6 ng/ml</td>
<td>0.003–0.75 ng/ml</td>
<td>0.01–23.0 ng/ml</td>
<td>0.02–24.0 ng/ml</td>
</tr>
</tbody>
</table>

processes (Zuvela et al., 2013). These manual assays have now been complemented by two fully automated assays (Dennis et al., 2014; Gassner and Jung 2014), with initial performance characteristics suggesting significantly more robust assays and lower inter-laboratory variation (Table 1) although there are no independent studies using either assay at present.

Additionally, confidence in the measurement of AMH has been shaken due to the recent field safety notices for the Beckman Coulter Gen II assay and the issue of complement interference (MHRA, 2013). Notably these issues were not observed in the original DSL assay nor the pre-release versions of the Gen II assay, both of which utilized the same antibodies (Wallace et al., 2011; Nelson et al., 2011a, b), suggesting that changing manufacturing processes may have been responsible. Other AMH assays are not affected by this issue (Su et al., 2014). A modified protocol has recently been externally validated as overcoming the interference observed in fresh samples (Han et al., 2014) although this is inconvenient to users. In contrast, as complement degrades with storage, interference in stored samples has been shown to be negligible (Welsh et al., 2014). Fortunately this interference issue is assay plate specific, as although the same antibodies are incorporated in the automated Elecsys® assay developed by Roche Diagnostics and the Access assay developed by Beckman Coulter, both the capture and signal generating antibody are in solution and do not bind complement (Dennis et al., 2014; Gassner and Jung 2014).

Potentially the most frustrating issue with measurement of AMH is the lack of an international standard developed in accordance with the International Federation of Clinical Chemistry. Given the increasing availability of a number of AMH assays and their widespread adoption into clinical practice there is an urgent need for an international human standard preparation to allow external calibration of AMH assays and standardize reporting and clinical interpretation. Figure 1 shows the Passing-Bablok regression analyses between AMH concentrations measured with the use of manual assays (Beckman AMH Gen II ELISA and Ansh Labs ultra-sensitive AMH/MIS ELISA) versus automated assay (Elecsys® AMH). At present the differences in calibration mean that AMH values by Gen II assay are ~20% higher and by Ansh assay ~30% higher than the new automated Elecsys® assay (Gassner and Jung 2014). This means that clinicians need to interpret the AMH result in an assay-specific manner, with adjustment for assay manufacturer not recommended for clinical practice. Furthermore the lack of an international standard impairs external assessment of commercial assay performance over time and between manufacturing lots.

Inter-individual variation

Concomitant with the decline in the rate of follicular recruitment observed with age in adult women, circulating AMH concentrations progressively decline with advancing age, reaching undetectable levels ~5 years prior to the cessation of menses (Sowers et al., 2008; Freeman et al., 2012). Several groups have modelled the age-related decline of AMH in large population cohorts, but all exhibit wide confidence intervals (CI) suggesting that for a given age AMH values in both normal and infertility populations can vary substantially (Almog et al., 2011b; Kelsey et al., 2011; Nelson et al., 2011a, b, 2014; Bentzen et al., 2013). This is not surprising as primordial follicle counts and follicular activity similarly vary substantially between individuals with a one hundred fold difference in primordial follicle numbers observed in healthy women of the same age (Wallace and Kelsey 2010).

Ethnicity has been associated with altered age-specific levels of AMH, with women of Chinese, Black African, Hispanic and South Asian descent reported as having a lower AMH at a given age compared with Caucasian women (Bleil et al., 2014; Iglesias et al., 2014). Whether this ethnic disparity reflects accelerated ovarian ageing, inherent differences in follicular endowment or recruitment and/or differences in AMH secretion per follicle is unclear. Clarification may be achieved by analyses of histological...
Biomarkers of ovarian response

701

Figure 1 Passing-Bablok regression analysis between anti-Müllerian hormone (AMH) concentrations measured with Beckman AMH Gen II enzyme-linked immunosorbent assay (ELISA) (top panel) and Ansh Labs ultrasensitive AMH/MIS ELISA (bottom panel) versus Elecsys AMH (reproduced from Gassner and Jung, 2014). The dotted line indicates equivalence, the solid line shows the actual regression.

ovarian specimens from different ethnic groups combined with assessment of follicular AMH secretion. The significance of altered AMH secretion is potentially substantial and, if proven, ethnicity-specific cut-points may be required in defining expected poor responders (Ferraretti et al., 2011) and high responders (Nelson et al., 2013).

Obesity is not thought to be associated with AMH levels. Initial cross-sectional data in 1896 non polycystic ovary syndrome (PCOS) infertile women demonstrated a weak negative correlation between AMH levels and BMI ($r = -0.064$); however, the authors did not adjust for age which would be expected to confound this association (Cui et al., 2014). This negative association was not replicated in a population-based study of 2320 premenopausal women with BMI, central adiposity and age-specific AMH percentiles (Dolleman et al., 2013). In an analysis of 1308 adolescent 15-year-olds with detailed DXA determined fat mass there was no association of AMH with fat mass or BMI (Anderson et al., 2013) and longitudinal analysis of women during weight loss did not demonstrate an alteration in AMH concentrations (Thomson et al., 2009; Vosnakis et al., 2012). With respect to other lifestyle determinants, cross-sectional data suggest that current smoking is independently associated with lower age-adjusted aggregated levels of AMH (Freour et al., 2008; Dolleman et al., 2013).

Comorbidities are increasingly recognized as being associated with altered AMH concentrations. The most established of these is PCOS, which is associated with substantially greater levels of AMH, to the extent that AMH cut-offs with optimal sensitivity and specificity have been suggested for the diagnosis of the syndrome (Dewailly et al., 2011; Iliodromitou et al., 2013; Lauritsen et al., 2014). For other medical disorders the data are more limited. Adolescent girls with type 1 diabetes have higher AMH levels than controls (Codner et al., 2011), but in adult life they are lower suggesting altered follicular dynamics—certainly women with type 1 diabetes go through the menopause earlier than healthy controls (Soto et al., 2009). In contrast in a study of 72 women with recent-onset rheumatoid arthritis no alteration of AMH was observed (Brouwer et al., 2013). More recently, it has been suggested that AMH production and/or follicular dynamics are altered acutely in response to being unwell. Young girls diagnosed with haematological or other childhood cancers exhibited decreased concentrations of AMH compared with their healthy peers at the time of initial diagnosis (van Dorp et al., 2014). In this cohort of 208 girls with newly diagnosed cancer AMH was also negatively associated with markers of general health including body temperature, C-reactive protein and anaemia. Similarly, lower AMH levels have been reported in adults with breast cancer, lymphoma and acute onset Crohn’s disease than in healthy (though infertile) controls (Lawrenz et al., 2012; Şenates et al., 2013; Su et al., 2013a). These data thus indicate the importance of general health in ovarian function, and consequently in interpreting tests of ovarian follicular activity.

In contrast to initial conclusions from cross-sectional studies it is now clear from prospective longitudinal studies that the endocrine environment which influences follicular activation and development also impacts on AMH concentrations. Pregnancy, GnRH analogues and combined hormonal contraceptives (irrespective of whether they are oral, transdermal or vaginal) are now all known to reduce AMH concentrations (Anderson et al., 2006; Nelson et al., 2010; Hagen et al., 2012; Kallio et al., 2013; Su et al., 2013b). This is likely to reflect suppression of endogenous gonadotrophin secretion and altered antral follicular development. Women would still be expected to respond to exogenous gonadotrophins as predicted by their AMH value; although there are limited data regarding this, it is consistent with the finding that fewer oocytes are obtained from women with a range of malignancies undergoing IVF prior to specific treatment (Friedler et al., 2012).

Although all the above factors may contribute to large age-specific variation in AMH, they do not seem to interfere substantially with the consistent robust associations with oocyte yield (La Marca and Sunkara 2014). This reflects that while AMH is expressed by granulosa cells from the initiation of follicle growth, expression is near-absent in cells from the final pre-ovulatory stages of development. In normal women it has been estimated that 60% of serum AMH is derived from follicles 5–8 mm in diameter (Jeppesen et al., 2006). This reflects that while AMH is expressed by granulosa cells from the initiation of follicle growth, expression is near-absent in cells from the final pre-ovulatory stages of development. In normal women it has been estimated that 60% of serum AMH is derived from follicles 5–8 mm in diameter (Jeppesen et al., 2006). Similar results were found by others (Hehenkamp et al., 2006; Tsepelidis et al., 2007; Streuli et al., 2008) although secondary analysis revealed that younger women (thus more likely to have higher AMH) had significantly larger intra-individual variation in AMH levels than older
women, with 17 out of 22 women under 38 years showing a variation in AMH concentration >0.5 ng/ml within one cycle (Overbeek et al., 2012). Evidence of statistically significant differences in mean values across the cycle can be misleading in studies aiming to assess the variation of AMH at the individual level. Analysis of the true intra-individual cycle variation indicated that the intraclass correlation coefficient (ICC) was 0.96 indicating that the between-subject variation was responsible for the larger proportion of the observed cyclical variation and only 4% of the variation was true within-subject variation related with the phase of the cycle (Deb et al., 2013).

Studies using the Gen II assay have confirmed cyclic variation with higher AMH in the late follicular phase (Hadlow et al., 2013), more evident in younger women with higher mean AMH levels (Kissell et al., 2014, Randolph et al., 2014). Figure 2 demonstrates the intra-cycle variation of AMH levels stratified by age categories in 259 women. Although the variation of AMH within a cycle was statistically significant, it had minimal impact on clinical performance and was not large enough to warrant a shift in clinical practice towards timing AMH measurement (Kissell et al., 2014).

To date, one study has addressed the circadian variation in AMH in a cohort of 19 women (Bungum et al., 2011). AMH (measured with the IOT assay) was lowest in the early morning hours (4 and 6 a.m.) with a maximum mean difference from its zenith values of 1.9 pmol/l (10.6%). AMH was relatively stable during daytime, when venepuncture is routinely performed, thus this result, though of interest, is not of clinical relevance. In contrast, the same study demonstrated that FSH, still used as a marker of ovarian response in some clinics, and other ovarian-derived hormones (estradiol and progesterone) exhibited substantial circadian fluctuation even during daytime (Bungum et al., 2011). Collectively, the data suggest that although AMH can vary across the menstrual cycle this variability may be primarily of potential value in detailed analysis of follicle growth patterns (Baerwald et al., 2012) and is not large enough to warrant restricting AMH measurements to a specific day or phase of the menstrual cycle.

The between cycle variability of AMH has been evaluated in studies using either IOT or DSL assay, but not the AMH Gen II assay. However, the results do not indicate assay specific variation. AMH, measured over 3 consecutive cycles, had an ICC of 0.89, which was significantly higher than that of FSH, inhibin B or AFC (0.55, 0.76, 0.73, respectively) (Fanchin et al., 2005) indicating lower variability. Age-adjusted ICC for AMH across four consecutive cycles was found to be 0.89 (95% CI, 0.84–0.94), indicating that only 11% of the inter-cycle variability was attributable to intra-individual fluctuation (van Dusseldorp et al., 2010); the ICC of AFC determined at the same time points was significantly lower (0.71, 95% CI 0.63–0.77). These analyses suggest that repeat measurements of AMH during subsequent cycles are not necessary for accurate patient assessment.

**Antral follicle count**

The developmental pathway from primordial follicle to ovulation is associated with approximately a 500-fold increase in follicular diameter (Charleston et al., 2007). Primordial follicles have a diameter of ~30 μm, and thus cannot be visualized; the development of the fluid-filled antrum provides the necessary physical structure to give a change in ultrasound reflectivity and thus allow potential detection, although this occurs at sub-millimetre diameters (Gougeon 1986) thus the smallest antral follicles cannot be visualized by current technology. Antral follicles in the range 2–10 mm can readily be counted on transvaginal sonography (TVS) to quantify an AFC and thereby predict ovarian response (Broekmans et al., 2006), although some clinics use a more limited range due to the increased variability in number of the larger follicles (Deb et al., 2013).

**Factors that can influence AFC**

**Technical issues**

The theoretical advantage of AFC over a biochemical marker is that TVS is available in any reproductive clinic; hence AFC can be readily performed and provides immediate results. That the wide availability of ultrasound may have compounded the technical issues has not been fully appreciated. Inter-observer and intra-observer variability in AFC determination have been robustly analysed (Hansen et al., 2003; Deb et al., 2009), illustrating the key limitation of this biomarker as currently performed. The commonly used 2-dimensional (2-D) technique in estimating the AFC has wide limits of agreement varying from +8 to −7 follicles when two consecutive measures are performed by the same operator or +7 to −5 follicles with two different experienced operators (Deb et al., 2009). This is sufficient to alter clinical management at an individual level and can introduce significant measure bias when pooling data in clinical research (Arce et al., 2013b). The reproducibility of the test improves only modestly when 3-D technique and offline analysis of the stored images is performed (Merce et al., 2005; Jayaprakasan et al., 2007, 2008; Deb et al., 2009), with additional analysis time and expense, increased workload and loss of the benefit of immediacy. The limits of agreement between consecutive measurements of AFC become significantly narrower when automatic analysis and counting or post-processing are implemented (Deb et al., 2009); however, the drawbacks of longer offline analysis persist and this approach has not been widely adopted. In addition, the validity of automated analysis is questionable given that it only counts approximately one-third of the antral...
Biomarkers of ovarian response

follicles measured with manual or the post-processing techniques (Deb et al., 2009).

A consensus statement was expected to resolve the issues of the large variability in AFC measurement by describing in detail the optimal technique with the use of the appropriate ultrasound probe in carefully selected patients (Broekmans et al., 2010). However, this report excluded women with previous ovarian surgery, ovarian endometriosis and women with a single ovary or irregular cycles, thereby excluding a significant proportion of the patients seen in a fertility clinic. Furthermore, the technical settings of ultrasound, such as depth, gain and focus, were not discussed (Broekmans et al., 2010). This contrasts with the use of ultrasound-based biomarkers such as nuchal translucency in prenatal diagnosis, which continuously undergo rigorous external quality assessment to ensure homogeneity and high accuracy in measurement by certified clinicians. An annual licensing arrangement run by the Fetal Medicine Foundation is undertaken for nuchal translucency, with retraining and removal of individual sonographers from their register the penalty for non-compliance. The lack of this standardization for AFC measurement may underlie its limited transportability across different operators, sites and settings. Potentially this issue could be addressed by international organizations, such as the European Society of Human Reproduction and Embryology, with ongoing quality control and annual renewal of licensing of AFC measurement subject to confirmation of adequate quality images. The degree of variation observed even in the research setting when the technique is as standardized as possible between centres (Arce et al., 2013b) illustrates the difficulties that need to be overcome.

The steady improvement in ultrasound resolution over the last decade has led to a recent re-evaluation of the AFC threshold for diagnosing polycystic ovarian morphology, now suggested to increase from ≥12 to ≥25 follicles (Dewailly et al., 2014b). That the upper limit for a normal AFC has more than doubled within such a short time frame is not dissimilar to the issues faced with the AMH assay and lack of standardization. The previous threshold of 12 follicles with currently available high resolution ultrasound would classify an expected ‘normal’ responder as ‘high’ responder with inappropriate selection of AFC stratified hyperstimulation protocols, and suboptimal stimulation for that patient. Likewise, the cut-off values suggested in studies for predicting poor response have evolved substantially from AFC<3 in 1998 (Chang et al., 1998) to <12 in 2009 (Melo et al., 2009). Thus the current suggested AFC thresholds of expected poor or normal ovarian response (Nelson et al., 2007, 2009; Ferraretti et al., 2011; Nelson 2013) may not be transferable to future ultrasound machines which will inevitably have higher resolution. This issue is also illustrated by AFC determination by magnetic resonance imaging, which gives significantly higher values than when measured by ultrasound (Leonhardt et al., 2014a, b). Inevitably AFC thresholds for clinical practice will always be subject to lagging behind the resolution of the available technology thereby adversely impacting on its potential role as a globally applicable biomarker.

With regard to patient acceptability of TVS we are not aware of work to date specifically targeting the views of infertility patients. Studies have reported that patients regard transvaginal scans as uncomfortable procedures but they are willing to undergo the procedure if recommended, while others have reported that only 10% of patients find the procedure embarrassing, stressful or uncomfortable (Dutta and Economides 2003, Basama et al., 2004). Additionally, transvaginal ultrasound provides a wealth of useful clinical data above that of just AFC measurement, justifying its place as a pivotal investigation for infertility patients (Kelly et al., 2001).

Inter-individual variation

Factors affecting AFC have been understudied relative to AMH but are likely to be similar given that they both reflect similar stages in the highly dynamic processes of follicular activity. In contrast to the large population cohorts for analysis of AMH, assessment of the relationship of AFC with age has only been examined in relatively small sample sizes (Broekmans et al., 2004; Almog et al., 2011a; La Marca et al., 2011; Wiweko et al., 2013). Although all confirm an age-related decline in AFC, they also recognize the substantial variation present at a given age. A single cross-sectional study suggested ethnic differences, with the average age-specific AFC in Indian women lower than in Caucasians. However the study did not provide 95% CIs for each ethnic specific regression line and was limited by its sample size (n = 229 Caucasians, n = 236 Indian women) (Iglesias et al., 2014). Smokers were reported as having a lower AFC compared with non-smokers of similar age (Freour et al., 2012). It is unclear whether this is a result of accelerated depletion of the primordial pool or modified follicular recruitment among smokers. The former mechanism of the effect of smoking on the ovaries has been suggested in animal models (Tuttle et al., 2009; Gannon et al., 2012) and may also be relevant in human fetal ovaries (Anderson et al., 2014) and linked with the increased risk of earlier onset of menopause among current smokers (Wellons et al., 2013).

As with AMH, AFC is reduced in cancer patients at the time of diagnosis (Ebbel et al., 2011). It is also now clear that AFC behaves similarly to AMH in response to exogenous hormones (Hansen et al., 2003). Gonadotrophin suppression caused by the contraceptive pill decreased the number of antral follicles, in particular of those measuring >6 mm in diameter (Deb et al., 2012). In line with this, a cross-sectional study showed that women taking the contraceptive pill had persistently lower AFC compared with women of the same age with natural cycles (Bentzen et al., 2012).

Confirmation of the effect size of these factors on AFC would be useful, but is unlikely to have an impact on the clinical application of AFC for prediction of ovarian response given that these factors are not modified prior to ovarian stimulation.

Intra-individual variation

AFC exhibits significant variation within and across consecutive cycles (Hansen et al., 2003; Etter et al., 2005; van Disseldorp et al., 2010; Deb et al., 2013). Assessment of the ICC of AFC showed that it had a modest ICC of 0.71 between two cycles and of 0.69 within one cycle, substantially worse than for AMH; 0.89 and 0.87, respectively (van Disseldorp et al., 2010). The source of this intra-cycle variability appears to be the variation in the number of the larger follicles (6–10 mm in diameter) (Deb et al., 2013). Given these concerns the consensus statement suggested that AFC should be performed from Day 2 to Day 4 of an index cycle (Broekmans et al., 2010). This recommendation is a significant limitation and inconvenience to both patient and clinic, and does not apply to women with irregular cycles.

In conclusion, despite its ready availability in every reproductive clinic, AFC is most accurately applied to well selected patients, has limited flexibility in relation to the phase of the cycle and exhibits substantial

Inter-individual variation

Factors affecting AFC have been understudied relative to AMH but are likely to be similar given that they both reflect similar stages in the highly dynamic processes of follicular activity. In contrast to the large population cohorts for analysis of AMH, assessment of the relationship of AFC with age has only been examined in relatively small sample sizes (Broekmans et al., 2004; Almog et al., 2011a; La Marca et al., 2011; Wiweko et al., 2013). Although all confirm an age-related decline in AFC, they also recognize the substantial variation present at a given age. A single cross-sectional study suggested ethnic differences, with the average age-specific AFC in Indian women lower than in Caucasians. However the study did not provide 95% CIs for each ethnic specific regression line and was limited by its sample size (n = 229 Caucasians, n = 236 Indian women) (Iglesias et al., 2014). Smokers were reported as having a lower AFC compared with non-smokers of similar age (Freour et al., 2012). It is unclear whether this is a result of accelerated depletion of the primordial pool or modified follicular recruitment among smokers. The former mechanism of the effect of smoking on the ovaries has been suggested in animal models (Tuttle et al., 2009; Gannon et al., 2012) and may also be relevant in human fetal ovaries (Anderson et al., 2014) and linked with the increased risk of earlier onset of menopause among current smokers (Wellons et al., 2013).

As with AMH, AFC is reduced in cancer patients at the time of diagnosis (Ebbel et al., 2011). It is also now clear that AFC behaves similarly to AMH in response to exogenous hormones (Hansen et al., 2003). Gonadotrophin suppression caused by the contraceptive pill decreased the number of antral follicles, in particular of those measuring ≥6 mm in diameter (Deb et al., 2012). In line with this, a cross-sectional study showed that women taking the contraceptive pill had persistently lower AFC compared with women of the same age with natural cycles (Bentzen et al., 2012).

Confirmation of the effect size of these factors on AFC would be useful, but is unlikely to have an impact on the clinical application of AFC for prediction of ovarian response given that these factors are not modified prior to ovarian stimulation.

Intra-individual variation

AFC exhibits significant variation within and across consecutive cycles (Hansen et al., 2003; Etter et al., 2005; van Disseldorp et al., 2010; Deb et al., 2013). Assessment of the ICC of AFC showed that it had a modest ICC of 0.71 between two cycles and of 0.69 within one cycle, substantially worse than for AMH; 0.89 and 0.87, respectively (van Disseldorp et al., 2010). The source of this intra-cycle variability appears to be the variation in the number of the larger follicles (6–10 mm in diameter) (Deb et al., 2013). Given these concerns the consensus statement suggested that AFC should be performed from Day 2 to Day 4 of an index cycle (Broekmans et al., 2010). This recommendation is a significant limitation and inconvenience to both patient and clinic, and does not apply to women with irregular cycles.

In conclusion, despite its ready availability in every reproductive clinic, AFC is most accurately applied to well selected patients, has limited flexibility in relation to the phase of the cycle and exhibits substantial
Comparison of AMH and AFC performance in predicting ovarian response

AMH and AFC have often been considered as interchangeable biomarkers for the prediction of ovarian response prior to commencement of ovarian stimulation. Acceptance of this is highlighted by their recent inclusion as alternative independent markers, combined with age, in the consensus statement on the definition of expected poor response (Ferraretti et al., 2011). At the other end of the ovarian response spectrum, despite excessive ovarian response not having an equivalent consensus statement definition, stratified stimulation algorithms with indicative starting dosing of gonadotrophins have been developed based on specific cut-offs of either single biomarker to avoid oocyte yields in excess of 15–20 oocytes (La Marca et al., 2012, 2013; Nelson 2013). Given that many centres may have had limited access to both biomarkers or exhibit preference for one or the other, there has been consideration of their potential overlap and their relative strengths.

Observational cohort data assessing predictive performance of AMH and AFC

In excess of 40 cohort studies and an IPD meta-analysis have examined the performance of AMH in the prediction of poor ovarian response (La Marca and Sunkara, 2014). Widely ranging threshold values for apparent optimal trade-off between sensitivity and specificity have been proposed. That the threshold values for AMH range from 0.1 ng/ml to 2.97 ng/ml primarily arises from the considerable heterogeneity introduced by the use of three different AMH assays, the inconsistencies in the definition of poor response and variable baseline characteristics and fertility potentials of the participants across the studies. Despite these limitations, the majority of the pooled studies reported that AMH has sensitivity >70% and specificity over 70% in predicting poor response in women undergoing fertility treatment (La Marca and Sunkara, 2014). Similarly, 22 cohort studies and one IPD meta-analysis have assessed the performance of AFC (La Marca and Sunkara, 2014); the performance characteristics of AFC also varied substantially among the studies, with threshold values ranging from 3 to 12, but the body of evidence suggests equivalent sensitivity and specificity to AMH in poor response prediction.

Single centre studies evaluating both biomarkers for the prediction of poor response have also generally not revealed significant differences in their performance (Van Rooij et al., 2002; Muttukrishna et al., 2005; Kwee et al., 2008; Jayaprakasan et al., 2010) although a minority of studies demonstrated significant superiority of one marker over the other; AFC over AMH (Mutlu et al., 2013) or AMH over AFC (Ficicioglu et al., 2006; McIlveen et al., 2007). A systematic review assessing the performance of each biomarker echoed the above findings summarizing that both biomarkers have equivalent receiver operating characteristic (ROC) curves in the prediction of poor response (Broer et al., 2009). This finding was replicated in a recent IPD analysis which demonstrated that the area under the curve (AUC) of the age-adjusted ROC curve for AMH in predicting poor response was 0.77 (95% CI: 0.70–0.83), practically identical to that of AFC (AUC 0.79, 95% CI: 0.73–0.85) (Broer et al., 2013b).

For excessive response the same issues have been observed; there have been in excess of 16 cohort studies and one IPD meta-analysis for AMH with a diverse range of threshold values and associated performance characteristics reported. For AFC there have been seven cohort studies and one IPD meta-analysis, but again no consensus on an overall threshold and anticipated performance. Cumulative analyses suggest comparable accuracy of AMH and AFC in predicting excessive ovarian response (Broer et al., 2011, 2013a; La Marca and Sunkara, 2014).

Collectively the above data provide apparent confirmation that both markers are equally effective at predicting poor and excessive ovarian response. However, observational cohort studies from individual clinics may have potentially inflated the performance of the association between exposure (ovarian reserve test) and outcome (response), particularly because the value of the test may have influenced the allocation of treatment and thus the outcome of interest (ovarian response) or through confounding, a known major limitation of observational studies. It is possible to reduce confounding in observational studies by restriction or matching, and in the statistical analysis by techniques such as stratification or multivariable analyses. These methods however require that the confounding variables are known and measured. Notably few of the single centre studies have undertaken this level of detailed analysis. In contrast a key strength of RCTs is that the randomization process allows the investigator to assume that not only known but also unknown potential confounders are distributed evenly among the treatment arms (Weinberg, 1993). Although the generalizability of RCTs can be limited due to the often stricter inclusion criteria and rigid protocols, RCTs are specifically designed to overcome the issues of differential confounding and selection bias between the treatment groups, making them strong candidates to examine the strength of association between exposures and outcomes of interest, and hence their widespread recognition as providing high-level evidence. The marked heterogeneity in reported threshold values and performance characteristics from the single centre studies implies that each individual centre would be required to develop its own thresholds. This does not have biological plausibility; there should not be marked heterogeneity in ovarian response of two biologically identical women treated in two different centres using an identical protocol. In the absence of such biological identity, we can assess how these models have performed in RCTs. Only one RCT has been specifically designed to compare AMH and AFC (with other markers) as predictive biomarkers (Andersen et al., 2011). However both have been included in four studies of protocols of ovarian stimulation for IVF (Andersen et al., 2011; Arce et al., 2013b; Nelson et al., 2014; Oehinger et al., 2013). The comparison of AMH and AFC in these is thus a secondary or post hoc analysis, and therefore potentially not as robust as if it were the primary analysis as it was one for one trial. Issues of study design are also relevant, particularly if AMH is measured centrally while AFC is derived locally which will inherently favour AMH. However the relative ease of standardization of hormone assay, with established quality control systems, such as the UK National External Quality Assessment Service (NEQUAS), compared with ultrasound analysis is an inherent potential advantage of AMH in determining clinically relevant cut-off values for widespread use as well as in multicentre research.

RCTs assessing predictive performance of AMH and AFC

Initial doubts about the equivalence of AFC and AMH in response prediction started appearing when the pharmaceutically sponsored international multicentre Xpect trial failed to show an independent association between AFC and the number of retrieved oocytes
In contrast, AMH was the only robust predictor of ovarian response in univariate and multivariate models. Similar findings were observed in both treatment arms, i.e., whether patients were randomly assigned to receive treatment with oral contraceptives prior to controlled ovarian stimulation or no pretreatment (Andersen et al., 2011). When poor or excessive response were assessed as dichotomous outcomes, AMH remained a significant predictive variable in each treatment arm, with no independent association observed for AFC. Notably the specific purpose of this trial was to identify factors capable of predicting ovarian response in patients undergoing their first treatment cycle with a daily dose of 200 IU recombinant FSH in a GnRH antagonist protocol, further strengthening the conclusion that AMH was the superior biomarker.

A subsequent pharmaceutically sponsored international multicentre RCT of two gonadotrophin preparations (MEGASET trial) added significant weight to the above findings. This study demonstrated a significant association between AMH and oocyte yield, number of blastocysts and cumulative live birth, but surprisingly it did not detect a significant univariate association between AFC and any of these outcomes (Arce et al., 2013b); consequently only AMH was associated with oocyte yield in multivariate models. Although the findings of the trial were initially criticized as potentially being attributable to marked operator variability across centres, secondary analysis demonstrated that there were only weak associations between AFC and oocyte yield within individual centres (Arce et al., 2013a). Retrospective analysis of another multicentre RCT (MERIT trial) comparing two different gonadotrophins in a long GnRH agonist protocol has also demonstrated that AMH had consistently greater association with ovarian response compared with AFC across different sites (Nelson et al., 2014). This site-specific analysis of the correlation of AFC and AMH with oocyte yield overcomes objections that AFC is an inherent limitation of AFC to predict ovarian response in a multicentre context, whereas AMH, when centrally analysed, is the more accurate biomarker performance in the prediction of ovarian response indicate the consistent data from several large scale RCTs assessing biomarker performance in the prediction of ovarian response indicate the inherent limitations of AFC to predict ovarian response in a multicentre context, whereas AMH, when centrally analysed, is the more accurate biomarker under those conditions.

### Conclusion

Significant changes have occurred in the measurement techniques for both AMH and AFC over the last decade such that the appropriate reference values for both biomarkers have changed substantially, and indeed further change is expected. Both reflect a very similar ovarian follicle population, thus, if perfectly measured, would be expected to
have similar value; this, supported by single site observational cohorts, underpins the classical viewpoint that these biomarkers exhibit equivalent performance characteristics for the prediction of ovarian response. However it appears likely that this equivalence has been overstated, inflated by study design, and emerging data from large scale multicentre RCTs indicate substantially better performance of AMH. International standardization of AMH combined with a robust automated assay is likely to enhance its status as the biomarker of ovarian response of choice. However the advantages of AFC will mean that it will continue to have an important role in clinical practice. The challenge for ultrasonography is to improve standardization of analysis to reduce the equipment- and observer-based aspects of variability. We therefore propose (Fig. 4) that AFC and AMH will have complementary roles in the pre-assessment of the infertile woman.

**Table II** Performance characteristics of prognostic models for ovarian response resulting from analyses of RCT data.

<table>
<thead>
<tr>
<th>Trials</th>
<th>Low ovarian response</th>
<th>High ovarian response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predictor variables</td>
<td>Performance characteristics</td>
</tr>
<tr>
<td>Xpect</td>
<td>AMH</td>
<td>AUC: 0.84</td>
</tr>
<tr>
<td></td>
<td>AMH &amp; smoking</td>
<td>AUC: 0.85</td>
</tr>
<tr>
<td>MEGASET</td>
<td>AMH</td>
<td>AUC: 0.78/0.90*</td>
</tr>
<tr>
<td></td>
<td>AFC</td>
<td>AUC: 0.67/0.74*</td>
</tr>
<tr>
<td>Engage^*</td>
<td>Age</td>
<td>AUC: 0.63</td>
</tr>
<tr>
<td></td>
<td>Age &amp; AFC</td>
<td>AUC: 0.75</td>
</tr>
<tr>
<td>Pursue</td>
<td>Age</td>
<td>AUC: 0.61</td>
</tr>
<tr>
<td></td>
<td>Age &amp; AMH</td>
<td>AUC: 0.87</td>
</tr>
<tr>
<td></td>
<td>Age &amp; AMH &amp; AFC</td>
<td>AUC: 0.88</td>
</tr>
<tr>
<td>MERIT</td>
<td>AMH</td>
<td>R^2: 0.29</td>
</tr>
<tr>
<td></td>
<td>AFC</td>
<td>R^2: 0.07</td>
</tr>
<tr>
<td></td>
<td>AMH &amp; AFC</td>
<td>R^2: 0.30</td>
</tr>
<tr>
<td>MEGASET</td>
<td>AMH</td>
<td>R^2: 0.23</td>
</tr>
<tr>
<td></td>
<td>AFC</td>
<td>R^2: 0.07</td>
</tr>
<tr>
<td></td>
<td>AMH &amp; AFC</td>
<td>R^2: 0.23</td>
</tr>
</tbody>
</table>

AUC: area under the curve of receiver operating characteristic curve, AFC: antral follicle count.

^*AMH was not measured.

^*Performance in each treatment arm.

**Figure 4** Relative contribution of ultrasound and AMH in the pre-IVF assessment of the female partner. We propose this as an optimal pathway for patient management reflecting the complementary roles of ultrasound and AMH in pretreatment assessment of women about to undergo IVF.
Authors’ roles

The review was developed from a presentation by S.M.N. at the 2nd Biomarker Meeting in Reproductive Medicine 24—26 April 2014 Valencia, Spain. S.I., R.A.A. and S.M.N. all contributed substantially to the design of the literature review. The systematic literature search was performed separately by S.I. and S.M.N., with S.I. and S.M.N. preparing the first draft. All authors edited and approved the final draft.

Funding

None declared.

Conflict of interest

R.A.A. has undertaken consultancy work and received speaker’s fees from Beckman Coulter, Roche Diagnostics and Ansh Laboratories. S.M.N. has undertaken consultancy work and received speaker’s fees from Beckman Coulter, Ferring Pharmaceuticals, Merck Serono, MSD and Roche Diagnostics.

References


Baker ML, Metcalfe SA, Hutson JM. Serum levels of mullerian inhibiting substance in boys from birth to 18 years, as determined by enzyme immunoassay. J Clin Endocrinol Metab 1990;70:11–15.


Beckman Coulter AMH Gen II ELISA package insert 2013.


Inn你就 asketh(p)leahome(at)com/wrporkal/axon/,%A1189320120605_Gb.pdf (May 2014, date last accessed).


Biomarkers of ovarian response


Reinares E, Calas A, Erdem ED, Yesil A, C¸oksel Matarı̈, Esin O, Altimir ME, Tuner I, Kadirı̈a O, Ov˘ı̈nc˘ AO. Serum anti-Mullerian hormone levels are lower in reproductive-age women with Crohn’s disease compared to healthy control women. J Crohns Colitis 2013;7:e29–e34.


