REVIEW

The accuracy of diagnostic tests for GH deficiency in adults: a systematic review and meta-analysis

Ahmad Hazem2, Mohamed B Elamin1, German Malaga4, Irina Bancos1,5, Yolanda Prevost4, Claudia Zeballos-Palacios4, Edgar R Velasquez4, Patricia J Erwin1, Neena Natt1,5, Victor M Montori1,5 and Mohammad Hassan Murad1,3

1Knowledge and Evaluation Research Unit, Mayo Clinic, 200 First Street SW, Rochester, Minnesota 55905, USA, 2Department of Internal Medicine, University of North Dakota, Fargo, North Dakota, USA, 3Division of Preventive, Occupational and Aerospace Medicine, Mayo Clinic, Rochester, Minnesota, USA, 4Universidad Peruana Cayetano Heredia, Lima, Peru and 5Division of Endocrinology, Diabetes, Metabolism, Nutrition, Mayo Clinic, Rochester, Minnesota, USA

(Correspondence should be addressed to M Hassan Murad at Knowledge and Evaluation Research Unit, Mayo Clinic; Email: murad.mohammad@mayo.edu)

Abstract

Context: The diagnostic accuracy of tests used to diagnose GH deficiency (GHD) in adults is unclear.
Objective: We conducted a systematic review and meta-analysis of studies that provided data on the available diagnostic tests.
Data sources: We searched electronic databases (MEDLINE, EMBASE, Cochrane CENTRAL, Web of Sciences, and Scopus) through April 2011.
Study selection: Review of reference lists and contact with experts identified additional candidate studies. Reviewers, working independently and in duplicate, determined study eligibility.
Data extraction: Reviewers, working independently and in duplicate, determined the methodological quality of studies and collected descriptive, quality, and outcome data.
Data synthesis: Twenty-three studies provided diagnostic accuracy data; none provided patient outcome data. Studies had fair methodological quality, used several reference standards, and included over 1100 patients. Several tests based on direct or indirect stimulation of GH release were associated with good diagnostic accuracy, although most were assessed in one or two studies decreasing the strength of inference due to small sample size. Serum levels of GH or IGF1 had low diagnostic accuracy. Pooled sensitivity and specificity of the two most commonly used stimulation tests were found to be 95 and 89% for the insulin tolerance test and 73 and 81% for the GHRH arginine test respectively. Meta-analytic estimates for accuracy were associated with substantial heterogeneity.
Conclusion: Several tests with reasonable diagnostic accuracy are available for the diagnosis of GHD in adults. The supporting evidence, however, is at high risk of bias (due to heterogeneity, methodological limitations, and imprecision).

European Journal of Endocrinology 165 841–849

Introduction

GH and the insulin-like growth factor 1 (IGF1) are involved in the regulation of somatic growth in children. In adults, this axis plays a role in maintaining normal body composition, skeletal mass, cardiovascular risk factors, and physical and physiological functioning (1). GH deficiency (GHD) has been related to insulin insensitivity, increased body fat, and decreased muscle mass; other studies have implicated that low levels of IGF1 are associated with increased risks of myocardial infarction, ischemic cardiac disease, and stroke (2). Thus, there has been interest in the clinical use of GH replacement in patients with GHD with the expectation that this agent could normalize body composition, enhance physical function, and prevent cardiovascular disease.

Among other indications, GH is approved for use in patients with hypothalamic–pituitary disease and GHD (3). GH is commonly used as an ‘anti-aging agent’ in patients without established deficit; in such instance, the evidence does not suggest substantial benefit (4). The therapeutic recommendation to use GH, therefore, rests in part on the ability to clinically diagnose patients who are truly deficient.

Expert reviews and consensus statements consider insulin-induced hypoglycemia (insulin tolerance test; ITT) as the reference standard for the diagnosis of this condition (3, 5). Alternative stimulation tests use GHRH + arginine, GHRH + GH-releasing peptide...
(GHRP), glucagon stimulation, and GHRP. To date, a summary of the available literature assessing the accuracy of the available tests for GHD in adult patients suspected of having the condition has not been conducted.

The aim of this systematic review and meta-analysis is to appraise and summarize the available evidence of the diagnostic accuracy of tests for GHD in adults.

Materials and methods

The protocol for this review is consistent with the available methodological guidelines for conducting diagnostic accuracy systematic reviews (6).

Eligibility criteria

We included cross-sectional and longitudinal studies that enrolled adult patients suspected of having GHD, i.e. clinicians had true diagnostic uncertainty about the patients’ GH status. Studies that enrolled patients with pre-existing diagnosis of GHD, as well as those that enrolled patients who were clearly healthy, were excluded because they would likely overestimate the diagnostic accuracy of the tests (7). Eligible studies assessed the accuracy of diagnostic tests by evaluating the true disease status in all enrolled participants according to a reference standard. We accepted any reference standard chosen by the included studies’ authors. When a reference standard was either unclear or not specified, we chose one of the tests of interest reported in the same study as a reference standard, usually opting for the ITT when available.

The ideal study design for our purpose was a randomized trial of testing treatment strategies in adult patients with suspected GH that evaluate a test for GHD in one arm versus no test or an alternative test. This design does not only yield accuracy data but also evidence about the utility of a particular management option based on a diagnostic test. However, we did not find any studies with such design.

Study identification and selection

An expert reference librarian (P J E) designed and conducted an electronic search strategy with input from study investigators with expertise in conducting systematic reviews (M H M and V M M). In order to identify eligible studies, we searched electronic databases (MEDLINE, EMBASE, Web of Science, Scopus, and citation search for key articles) through April 2011. The detailed search strategy is available on request. Furthermore, we sought references from experts in the field.

Two reviewers, working independently, considered the potential eligibility of each of the abstracts and titles that resulted from executing the search strategy. Following the same procedures, reviewers evaluated the full-text reports (all available versions of each study) for eligibility. Subsequently, disagreements were harmonized by consensus; if not possible, by arbitration (i.e. a third reviewer adjudicated the disagreement).

Data extraction

Reviewers working independently and in duplicate used a standardized web-based form to extract the description of study participants, comorbidities, and chosen reference standard and cutoff value.

To assess diagnostic accuracy, we used the cutoff values chosen by the study authors. When multiple or no cutoff values were reported, we tended to borrow a cutoff value from another study that described the same test. In the instances that a study was assessing a test that had not been described in the previous literature, we chose the cutoff value that yielded the highest diagnostic accuracy of the test.

Author contact

Letters were sent to the corresponding authors (or if not reported the first and last authors) of each of the 28 studies that reached data extraction level by electronic mail (regular mail when an e-mail was not obtainable). We asked authors to verify the data we extracted and to complete missing data that we could not identify in the published record.

Statistical analysis

We used Meta-DiSc Software for Meta-analysis for Screening and Diagnostic tests version 1.4. We used random effect meta-analyses to pool the sensitivities, specificities, likelihood ratios, and diagnostic odds ratio (DOR) and estimate the 95% confidence intervals for the outcomes.

The DOR of a test describes the ratio of the odds of a positive test result in patients with disease compared to patients without disease (8) and can be calculated as the ratio of the likelihood ratios for a positive and a negative test. Its main advantage lies in it being a single value that is indicative of a general test performance compared to the gold standard.

Summary receiver operating characteristic (ROC) curves facilitate visual assessment of the consistency of results across studies and the accuracy of the test, estimated by the area under the summary ROC curve, in differentiating between patients with GHD and patients without GHD. Unlike ROC curves in which individual data points represent different test cutoffs, each point in a summary ROC curve represents a study (9). In order to assess the inconsistency across studies, we have used the $I^2$ statistic, which represents the proportion of variability across studies that is not attributable to
chance. $I^2$ values of 25, 50, and 75% indicate low, moderate, and high heterogeneity respectively.

**Subgroup analyses**

A *priori* hypotheses to explain the potential heterogeneity across studies included choice of reference standard, history of childhood onset GHD, presence of other pituitary hormone deficiencies, presence of known pituitary disease (history of meningitis, brain trauma, subarachnoid hemorrhage, and head irradiation), and choice of threshold (data driven, assay driven, from previous study). We tested these hypotheses using a test for interaction (10) considering $P<0.05$ as significant. There were not enough studies to use meta-regression as a strategy to identify predictors of test accuracy and explain inconsistency in results across studies.

**Results**

**Study identification**

Initial search of the literature yielded 952 publications, of which 117 were potentially relevant to this review based on the titles and abstracts. After full-text review, we found 28 eligible studies. We excluded seven studies from analysis because they did not report sufficient data for meta-analysis or did not measure the outcomes of interest. Attempts had been made to contact the authors of these studies and were either unsuccessful or unfruitful (11–18).

We asked authors to verify the data we extracted and to complete missing data that we could not identify in the published record. We were able to establish contact with 17 (60%) of the authors of the 28 studies, of whom 14 (82%) responded either by providing missing data or by confirming the accurate representation of their studies.

**Table 1** Characteristics of included studies. The patient characteristic categories in the table are not mutually exclusive and the percentages do not add to 100%.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Population characteristics</th>
<th>Population description (% from total)</th>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y; ± S.D. or Range)</td>
<td>Size</td>
<td>GHD*</td>
<td>CGHD</td>
</tr>
<tr>
<td>(1) 36.4 ± 2.1</td>
<td>40</td>
<td>NR</td>
<td>47.5</td>
</tr>
<tr>
<td>(32) R: 18–75</td>
<td>49</td>
<td>4.1</td>
<td>NR</td>
</tr>
<tr>
<td>(33) M: 39.7, F: 40.3</td>
<td>47</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>(34) 52 (23–77)</td>
<td>34</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(35) R: 28–64</td>
<td>49</td>
<td>NR</td>
<td>0</td>
</tr>
<tr>
<td>(36) 65.8 (60–72)</td>
<td>18</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(37) 19.2 ± 0.2 (s.e.m.)</td>
<td>152</td>
<td>NR</td>
<td>100</td>
</tr>
<tr>
<td>(38) 24.3</td>
<td>11</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(39) 46 ± 16</td>
<td>142</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(40) 43.1 (21–59)</td>
<td>26</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(41) 24.5</td>
<td>108</td>
<td>NR</td>
<td>100</td>
</tr>
<tr>
<td>(42) 25.7 ± 5.7</td>
<td>69</td>
<td>NR</td>
<td>100</td>
</tr>
<tr>
<td>(43) R: 18–59</td>
<td>36</td>
<td>0</td>
<td>19.44</td>
</tr>
<tr>
<td>(44) 53.32 (31–76)</td>
<td>19</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(45) R: 19–66</td>
<td>20</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(46) 23.2 ± 1.4</td>
<td>22</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(47) 40.2 ± 12.1</td>
<td>21</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(48) 41 ± 13</td>
<td>30</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>(49) 53 ± 2 (s.e.m.)</td>
<td>35</td>
<td>NR</td>
<td>0</td>
</tr>
<tr>
<td>(50) 22.5</td>
<td>6</td>
<td>NR</td>
<td>100</td>
</tr>
<tr>
<td>(51) 35.76</td>
<td>138</td>
<td>NR</td>
<td>0</td>
</tr>
</tbody>
</table>

GHD, GH deficiency; NR, not reported; ITT, insulin tolerance test; PD, pyridostigmine; AST, arginine stimulation test; IGF1, insulin-like growth factor 1, CGHD, childhood onset GHD; PD, pituitary disease; PHD, pituitary hormone deficiencies; COM, comorbidities. * presumed GHD, otherwise healthy.
analyses explored were not associated with significant
decreased to some extent the overestimation of test
accuracy subgroup interactions (P > 0.05). However, the interaction tests are severely under-
powered, and therefore, the results may not be reliable (see Supplementary Table 1, see section on supple-
mentary data given at the end of this article).

Discussion

Summary of findings

We conducted a systematic review and meta-analysis to
evaluate the diagnostic accuracy of tests for GHD. Twenty-three studies provided diagnostic accuracy
data; none provided patient outcome data. Studies had
fair methodological quality (assessed using the QUADAS
checklist), used several reference standards, and
included over 1100 patients. Tests involving GH
stimulation had good diagnostic accuracy (DOR > 50).
Heterogeneity was significant in most analyses. The
only significant subgroup interaction was found with
relation to serum IGF1 level test and the presence or
lack, thereof, of known pituitary disease.

Limitations and strengths

Inferences presented in this review are limited by the
heterogeneity of gold standard definition and cutoffs and
in the prevalence of GHD across studies. Imprecision
was present in several analyses and lowers the quality of
the evidence. This was evident for GHRP6, GHRH,
acipimox+GHRH, GHRH+GHRP6, GHRH+GHRP2,
hexarelin stimulation, GHRH+clonidine, arginine
stimulation, GHRH+pyridostigmine, GHRH+SMS,
and glucagon stimulation test that were evaluated in
only one or two studies leading to estimates with very
wide confidence intervals. Clinicians need to interpret
imprecise results with caution and have lower confi-
dence when applying these results at the point of care.
Imprecision has likely affected subgroup analyses as
well, and a true interaction might have existed but was
undetected by the underpowered analyses.

We were unable to evaluate the presence of
publication bias because the studies included in each
analysis of a diagnostic test were <20 and hetero-
geneity was significant. Under these conditions, all the
available methods (funnel plots symmetry, Egger’s
regression, etc.) are unreliable and function under the
wrong assumptions (19).

This review successfully summarizes all available
evidence in a systematic (i.e.: protocol-driven) approach
leading to reproducible results (20). By restricting our
criteria to include studies that had exclusively enrolled
patients with a diagnostic uncertainty (21), we have
decreased to some extent the overestimation of test
accuracy that can otherwise result, namely, in studies
that enroll patients who have a pre-established
diagnosis of GHD and compare them to healthy adults.
Table 2 Meta-analyses results – summary of diagnostic accuracy measures.

<table>
<thead>
<tr>
<th>Test</th>
<th>No. of studies</th>
<th>Gold standard</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Likelihood ratio of a positive test</th>
<th>Likelihood ratio of a negative test</th>
<th>DOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHRP6 Serum IGF1 levels</td>
<td>2</td>
<td>ITT, 2: GHRH + arginine, 1: clonidine, 1: GHRH + PD, 1: AST, 1: concordance between ITT, AST and GHRH, or serum IGF1 levels in the absence of concordance, 1: glucagon</td>
<td>0.85 (0.70–0.94) 82.5</td>
<td>0.97 (0.82–0.99) NA</td>
<td>16.93 (3.54–80.94) NA</td>
<td>0.16 (0.05–0.58) NA</td>
<td>109.58 (15.88–755.96) NA</td>
</tr>
<tr>
<td>GHRH + arginine Acipimox + GHRH</td>
<td>1</td>
<td>ITT</td>
<td>0.72 (0.68–0.77) 90.3</td>
<td>0.63 (0.57–0.67) 84.1</td>
<td>2.28 (1.59–3.28) 63.5</td>
<td>0.44 (0.27–0.72) 90.7</td>
<td>8.43 (4.42–16.09) 36.5</td>
</tr>
<tr>
<td>GHRH + GHRP6</td>
<td>2</td>
<td>AST, ITT</td>
<td>0.70 (0.60–0.79) 82.5</td>
<td>0.82 (0.71–0.90) 77.9</td>
<td>2.41 (0.81–7.15) 74.6</td>
<td>0.64 (0.41–0.99) 55</td>
<td>5.62 (2.16–14.63) 0</td>
</tr>
<tr>
<td>Hexarelin stimulation GHRH + GHRP2</td>
<td>1</td>
<td>ITT</td>
<td>0.79 (0.59–0.92) NA</td>
<td>1.00 (0.84–1.00) NA</td>
<td>16.96 (2.45–117.32) NA</td>
<td>0.29 (0.14–0.61) NA</td>
<td>64.2 (6.69–625.77) NA</td>
</tr>
<tr>
<td>GHRH + clonidine Arginine</td>
<td>1</td>
<td>Concordance between ITT, AST and GHRH, or serum IGF1 levels in the absence of concordance</td>
<td>1.00 (0.84–1.00) NA</td>
<td>1.00 (0.66–1.00) NA</td>
<td>19.55 (1.30–291.53) NA</td>
<td>0.02 (0.00–0.37) NA</td>
<td>817.00 (15.06–44 322.80) NA</td>
</tr>
<tr>
<td>GHRH + PD</td>
<td>2</td>
<td>ITT</td>
<td>0.93 (0.83–0.98) NA</td>
<td>0.75 (0.48–0.93) NA</td>
<td>3.59 (0.48–27.12) NA</td>
<td>0.10 (0.03–0.34) NA</td>
<td>35.01 (2.12–577.38) NA</td>
</tr>
<tr>
<td>Serum GH levels GHRH</td>
<td>3</td>
<td>1: ITT, 1: GHRH + PD, 1: AST</td>
<td>0.23 (0.19–0.54) 0</td>
<td>0.87 (0.64–0.98) 46.2</td>
<td>1.60 (0.33–7.76) 6.8</td>
<td>0.85 (0.58–1.25) 19.2</td>
<td>2.09 (0.29–15.07) 10.3</td>
</tr>
<tr>
<td>GHRH + SMS</td>
<td>1</td>
<td>Concordance between ITT, AST and GHRH, or serum IGF1 levels in the absence of concordance</td>
<td>0.80 (0.66–0.90) NA</td>
<td>0.79 (0.49–0.95) NA</td>
<td>3.02 (0.71–12.77) NA</td>
<td>0.15 (0.00–8.48) NA</td>
<td>21.51 (0.27–1736.50) NA</td>
</tr>
<tr>
<td>GHRH + Glucagon stimulation test</td>
<td>1</td>
<td>ITT</td>
<td>0.95 (0.76–1.00) NA</td>
<td>0.79 (0.66–0.88) NA</td>
<td>4.44 (2.17–9.09) NA</td>
<td>0.06 (0.01–0.41) NA</td>
<td>73.33 (8.11–663.10) NA</td>
</tr>
<tr>
<td>ITT</td>
<td>3</td>
<td>1: Concordance between ITT, AST and GHRH, or serum IGF1 levels in the absence of concordance</td>
<td>0.95 (0.86–0.99) 30.4</td>
<td>0.89 (0.74–0.96) 0</td>
<td>6.54 (2.95–14.49) 0</td>
<td>0.08 (0.03–0.26) 0</td>
<td>106.86 (17.62–648.22) 0</td>
</tr>
</tbody>
</table>

GHRP, GH-releasing peptide; ITT, insulin tolerance test; AST, arginine stimulation test; SMS, somatostatin; IGF1, insulin-like growth factor 1; PD, pyridostigmine.
By using the DOR, we hope to have sufficiently compensated for inconsistencies across included studies, in terms of choice of cutoffs.

**Implications for practice and research**

Several clinical practice guidelines on the diagnosis and management of adult GHD recommend the use of provocative tests of GH secretion within an appropriate clinical context and also refer to published diagnostic cutoff criteria (3, 5, 22). Most of the studies included in this review analyzed the two most commonly recommended tests in these consensus statements, namely the ITT and the GHRH-arginine test, but only the former was reported to have good diagnostic accuracy based on the DOR.

From the clinical practice point of view, it is important to recognize that the diagnostic accuracy of provocative tests of GH secretion is hampered by a lack of standardization of GH assays (23). Thus, the same patient blood sample sent to different laboratories can yield different GH concentrations. This discrepancy between GH assays can be attributed to a number of factors including variable antibody recognition of GH isoforms, the use of different GH preparations for calibration, lack of consensus regarding units of measurement for reporting GH, and variability of unit conversion factors (24). It is, therefore, important for clinicians to know which GH analytic method their laboratory uses and whether diagnostic cutoff values for GH provocative tests have been adjusted for that particular assay. Patient characteristics can also influence accuracy of interpretation of GH provocation tests. Obese normal subjects have been shown to exhibit a blunted GH response to provocative stimuli to a degree that sometimes overlaps with the results obtained from severe GH-deficient adults (25). It has, therefore, been suggested that diagnostic cutoff criteria should be adjusted for body mass index to avoid the potential of a false-positive result. Fortunately, IGF1 levels are not low in obese normal patients, and hence, a positive GH stimulation test coupled with a low IGF1 in an obese patient in the appropriate clinical setting is often reassuring to the clinician that the diagnosis of GHD is accurate.

This review's finding of poor diagnostic accuracy of the IGF1 concentration in patients suspected of having GHD is in keeping with reports that IGF1 levels show considerable overlap between normal and GH-deficient adults. Hence, a normal IGF1 level does not rule out GHD. However, the presence of a low IGF1 level in patients with hypopituitarism associated with three or more pituitary hormone deficiencies is considered highly indicative of GHD (26, 27), and clinical guidelines suggest that provocative tests of GH stimulation are optional in such cases. Another confounder when using IGF1 assays is that some reference

---

**Figure 2** (A) ITT SROC curve. (B) Serum IGF1 SROC curve. (C) GHRH-arginine SROC curve. (D) Serum GH SROC curve.
data, Yuen et al. (28) have advocated the glucagon stimulation test as an alternative to the ITT until another recombinant form of GHRH or alternative GH secretagogues used in the studies analyzed by this review are also not available for clinical use in the US. This has now raised the question of which GH stimulation test should be used in clinical practice when the ITT is contraindicated or cannot be performed due to lack of required resources. Based on the review of available data, Yuen et al. (28) have advocated the glucagon stimulation test as an alternative to the ITT until another recombinant form of GHRH or alternative GH secretagogue becomes available (phase III studies of an oral ghrelin agonist are in process). The American Association of Clinical Endocrinologists has also recommended the glucagon stimulation test in place of the GHRH-AST (22). The advantages of the glucagon test are mainly its wide availability in the United States, the ease with which the 4 h test can be completed, and its high relative tolerability. We found one recently published study that assessed the diagnostic accuracy of glucagon stimulation test and reported a high sensitivity and specificity and we calculated a DOR of over 70 (29). Two other studies assessed the accuracy of the glucagon stimulation test but had to be excluded from our meta-analysis due to the fact that the studies enrolled patients who had a pre-established diagnosis of GHD (30, 31). The two studies reported sensitivity and specificity ranges of 97–100 and 88–100% respectively.

Conclusion

Several GH stimulation tests with fairly good diagnostic accuracy are available for the diagnosis of GHD in adults. The supporting evidence, however, is indirect (as is not directly informative of patient outcomes) and at high risk of bias.

Supplementary data

This is linked to the online version of the paper at http://dx.doi.org/10.1530/EJE-11-0476.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This review was commissioned and funded by a contract for the Endocrine Society (Grant no.: not applicable).

References

11 Aimaretti G, Baffoni C, Bellone S, Di Vito L, Corneli G, Arvat E, Benso L, Camanni F & Ghigo E. Retesting young adults with...
38 Darzy KH, Thorner MO & Shalet SM. Cranially irradiated adult cancer survivors may have normal spontaneous GH secretion in the presence of discordant peak GH responses to stimulation tests (compensated GH deficiency). Clinical Endocrinology 2009 70 287–293. (doi:10.1111/j.1365-2265.2008.03159.x)
41 Juul A, Kastrup KW & Skakkebøe NE. Growth hormone (GH) provocative retesting of 108 young adults with childhood-onset GH deficiency and the diagnostic value of insulini-like


45 Petersenn S, Jung R & Beil FU. Diagnosis of growth hormone deficiency in adults by testing with GHRP-6 alone or in combination with GHRH: comparison with the insulin tolerance test. European Journal of Endocrinology 2002 146 667–672. (doi:10.1530/eje.0.1460667)


Received 29 May 2011
Revised version received 18 August 2011
Accepted 19 August 2011