



Published in final edited form as:

J Am Geriatr Soc. 2006 November ; 54(11): 1666–1673.

Androgen Treatment and Muscle Strength in Elderly Males: A Meta-Analysis

Kenneth J. Ottenbacher, Ph.D.¹, Margaret E. Ottenbacher, M.S., R.N.², Allison J. Ottenbacher, M.S.³, Ana Alfaro Acha, M.D.⁴, and Glenn V. Ostir, Ph.D.^{1,5}

¹Division of Rehabilitation Sciences, Sealy Center on Aging, University of Texas Medical Branch, Galveston, Texas

²Division of Geriatrics, University of Texas Medical Branch, Galveston, Texas

³School of Public Health, University of Texas Health Science Center, Houston, Texas

⁴ Division of Geriatric Medicine, Hospital Virgen del Valle, Complejo Hospitalario de Toledo, Spain

⁵ Department of Internal Medicine, University of Texas Medical Branch, Galveston, Texas

Abstract

Objectives—Research regarding the effectiveness of androgen treatment to increase muscle strength in older men is contradictory. We reviewed published, randomized trials examining the impact of androgen treatment on muscle strength.

Design—Systematic review using meta-analysis procedures.

Setting & Participants—We searched for trials in MEDLINE, EMBASE, CINAHL, and the Cochrane Register. Key words included testosterone, androgen, sarcopenia, muscle loss, aged, aging, elderly, older, geriatric, randomized controlled trials, and controlled clinical trials. 65 non-overlapping studies were found. Meta-analysis methods were used to evaluate the 11 randomized, double-blind trials.

Intervention—Testosterone or dihydrotestosterone (DHT) replacement therapy in healthy men 65 years and older.

Measurements—Tests of muscle strength.

Results—The studies included 38 statistical comparisons. The mean *g*-index adjusted for sample size was 0.53 (95% CI, 0.21-0.86). Sub-analyses revealed larger effects for measures of lower extremity ($g_i = 0.63$, 95% CI, 0.31-1.28) than for upper extremity muscle strength ($g_i = 0.47$, 95% CI, 0.12-0.84). A larger mean *g*-index was found for injection ($g_i = 0.95$, 95% CI, 0.30-1.58) versus topical ($g_i = 0.26$, 95% CI, 0.08-0.42) or oral ($g_i = -0.21$, 95% CI, -1.40-1.02) administration of testosterone/DHT. Effect sizes were related to study characteristics such as subject attrition and design quality ratings. Sensitivity analyses revealed the elimination of one study reduced the mean *g*-index from 0.53 to 0.23.

Conclusion—The results suggest that testosterone/DHT therapy produced a moderate increase in muscle strength in men participating in 11 randomized trials. The mean effect size was influenced by one study.

Correspondence and reprints to: K. Ottenbacher University of Texas Medical Branch 301 University Blvd. Galveston, TX 77555-1137 kottenba@utmb.edu (409) 747-1637 (O), (409) 747-1638 (F). Alternate Corresponding Author: G. Ostir University of Texas Medical Branch 301 University Blvd. Galveston, TX 77555-0460 gostir@utmb.edu (409) 747-1880 (O), (409) 772-8931 (F).

Sources of Support: This research was supported by the National Institute on Aging grant K02 AG01973 (K. J. Ottenbacher) and the National Center for Medical Rehabilitation Research grant K01 HD046682 (G.V. Ostir).

Keywords

Hormone; Testosterone; Exercise; Evaluation

INTRODUCTION

Hormone therapy and the use of androgens to improve health and maintain independent function in older adults has received substantial professional and public attention.¹⁻⁴ Recent reports from large randomized trials have raised concerns regarding the widespread use of hormone replacement therapy for postmenopausal women.^{5,6} Hormone replacement and androgen supplementation are also popular for older men. Testosterone prescriptions for men have increased by more than 500% in the past decade.²

Hormone replacement in men is a popular intervention because testosterone and other androgens are inexpensive and are known to have anabolic effects on muscle, fat, and bone.⁴ The rationale for the use of anabolic therapies to enhance physical function is based on the premise that hormone replacement therapy, through alterations in intramuscular gene expression, increases muscle mass, and that increased muscle mass translates into improved physical performance.⁷ Two recent reviews examined the pros and cons of androgens in treating male andropause.⁸ Asthana and colleagues reviewed the impact of testosterone therapy on skeletal muscle, bone density, cognition, erectile function, prostatic hyperplasia and cancer, insulin sensitivity, and cardiovascular function and concluded that previous studies have demonstrated modest gains in lean muscle mass and decreases in fat mass, “but inconsistent changes in muscle strength”.⁸ Gruenewald and Matsumoto⁹ reported a similar review focusing on older men (> 60 years) and concluded that testosterone's effects on strength and function were mixed.

A systematic review of interventions for sarcopenia and muscle weakness in adults by Borst concluded that “testosterone replacement therapy in elderly hypogonadal men produces only modest increases in muscle strength, which are observed in some studies and not others”.¹⁰ This review examined longitudinal studies and clinical trials in younger and older men. No attempt was made to determine the impact of study design or sample size on outcome, or to compute a standardized metric (effect size) between comparison groups.

Muscle strength is a key factor in maintaining functional independence in older adults. Decreased muscle strength is a risk factor for frailty and disability.^{11,12} A recent study by Sullivan and colleagues¹³ combined testosterone with progressive, resistance muscle-strength training in frail older men. All four groups received either low or high resistance exercise. Two groups received testosterone as well, and the other two received placebo. The authors concluded that the addition of testosterone produced greater muscle size and a trend toward increased muscle strength. This was a well-designed study, but the sample sizes in the groups were small (< 20). If testosterone therapy in men increases lean muscle mass, but does not produce improvements in strength, its therapeutic effectiveness maybe questioned, particularly in view of the potential risks related to prostate cancer and cardiovascular disease.^{3,7}

The purpose of the current study was to examine the ability of androgen therapy to increase muscle strength in older men. We conducted a meta-analysis of randomized-controlled trials involving men over the age of 65 and receiving treatment with androgens or placebo.

METHODS

Potentially, relevant clinical trials were identified through computerized and manual searches. We elected to include only randomized, double-blind trials that were published in refereed

journals. This decision was based on the assumption that the results from published randomized trials represented the highest level of evidence for guiding decision making in clinical practice.¹⁴ Computer-aided searches were conducted using the following databases: MEDLINE, EMBASE, CINAHL, and the Cochrane Register. The key words used in the search included testosterone, androgen, sarcopenia, muscle loss, aged, aging, elderly, older, geriatric, randomized controlled trials, and controlled clinical trials. Key words were used in various combinations to maximize search results. The reference lists from articles identified in the computer searches were examined to identify potentially relevant investigations. The method of citation tracking was used to further identify authors who published potentially relevant studies. Additional searches were conducted in the databases using the names of authors who published previous clinical trial articles on hormone therapy in men. The search covered articles published between 1980 and 2005. Approval of this meta-analysis was not required by the Institutional Review Board of the University of Texas Medical Branch, Galveston, Texas.

Decision Criteria

Inclusion criteria were pre-defined and included: 1) refereed journal publications (in English); 2) participants with a mean age of 65 years or greater; 3) randomized, clinical-trial design including comparisons between patients who received testosterone or dihydrotestosterone (DHT) and patients who received placebo; 4) intervention including the administration of testosterone or DHT, with specified method and dosage; 5) an operationally-defined measure of upper and/or lower extremity muscle strength; and 6) sufficient, statistical information reported to estimate a standardized mean difference (effect size).

A clinical trial had to report a comparison between at least two groups or conditions. In the majority of cases, the comparison was between a treatment group that received testosterone and a comparison group that was provided a placebo. In some trials where a within-subject or cross-over design was used, the comparison condition included the same patients who later served as the experimental group. All studies were coded based on whether the comparison was between or within subjects (see below).

Coding of Clinical Trials

With the boundaries of the review determined, the next step was to identify aspects of the trials related to patient outcomes. These variables fell into four general categories. The first category was patient characteristics, including information on the number of subjects participating in the trial and their mean age, educational level, race, etc. The second category incorporated information related to the independent and dependent variables and the design characteristics of the trial, including method of randomization, duration of intervention, method of testosterone/DHT administration (injection, topical or oral), and whether the outcome measure involved upper and/or lower extremity strength and how many times it was recorded. In studies where the dependent variable was recorded repeatedly, we used the measure that was closest to the termination of intervention to estimate the effect size (see description below).

The third category included aspects of the trial's outcome such as statistical test used, test value reported, accompanying probability level, and degrees of freedom associated with error. Design quality was coded based on criteria described by Lipsey and others.^{15,16} The coding form included thirteen questions related to design quality (e.g., hypotheses clearly presented, reliability of measures reported, etc.) Each question was rated yes, no, or information not provided. Yes was scored as 1 and no or missing was scored as 0. A total quality score was computed ranging from 0 to 13 with a higher score indicating a better design quality rating. In addition to the design quality rating, each study was coded based on whether or not the design and analyses were identified as intention-to-treat. Studies were coded regarding whether information on subject attrition was reported. Attrition was rated as 1 = 0 attrition, 2 = 1 to

10%, and 3 =>10% attrition. The final category, labeled retrieval characteristics, included the setting in which the trial was conducted and the year and source of publication.

Coding Reliability—The eleven studies were coded by two raters. The raters reviewed each of the 11 investigations and independently completed the coding sheet. The introduction and methods section of each article were coded prior to recording the results and without knowledge of the study findings or effect size estimates obtained for a given study. Interrater agreements for all items were computed using the intraclass correlation (ICC) approach for continuous variables and Kappa for categorical variables. The ICC values ranged from 0.77 to 1.00, and Kappa ranged from 0.56 to 1.00 for items that were used in subsequent analyses.

Quantifying Study Outcomes

Cohen¹⁷ and others¹⁸ have popularized procedures capable of uncovering systematic variation in the statistical outcomes of clinical trials. These procedures involve the calculation of a standardized mean difference (effect size) and the analysis of effect sizes in relation to study and design characteristics. The effect size measure used in this study was the g-index (g_i).¹⁸ The g-index gauges the difference between two group means in terms of their common (average) standard deviation. The effect size we used is the weighted mean of the unbiased individual effect sizes, calculated according to a random effects model.¹⁸ Each individual g-index was calculated by dividing the difference between the gains of the experimental and placebo groups by the estimated, pooled standard deviation index (S_{pooled}) of the baseline outcome measure in the treatment and placebo groups, where:

$$S_{pooled} = \sqrt{\frac{Sg_1^2(ng_1 - 1) + Sg_2^2(ng_2 - 1)}{(ng_1 - 1) + (ng_2 - 1)}}$$

with Sg_1 and Sg_2 = standard deviations for groups 1 and 2, and ng_1 and ng_2 = sample sizes for groups 1 and 2.

Effect sizes were adjusted according to study sample size using the Mantel-Haenzel weighting procedure.¹⁹ This approach involves estimating a weighting factor which is the inverse of the variance associated with each g-index estimate. If studies did not report specific statistical values required to compute the g-index, we contacted the authors to request raw data by email or phone. A minimum to two contacts were made to obtain raw data. If authors did not provide raw data, we assumed an effect size of 0.00.

Effect sizes were calculated using the Comprehensive Meta-Analysis software package (BioStat, Englewood, New Jersey). Standardized mean difference effect sizes (g_i) were computed from summary statistics such as means and standard deviations, *t*-tests and *F*-tests. In studies where a non-significant result was reported and no calculated statistical values were provided, an effect size of 0.00 was assumed, as noted above. Zero effect size values were assigned for eight of the 38 statistical comparisons. We used the estimated 0.00 effect size in determining the mean effect size across all comparisons. We did not use 0.00 standard errors or confidence intervals in determining means or average measures of variation.

A g-index was computed for each measure of muscle strength reported in the 11 studies. This meant that there were more g-indexes than there were studies, since several studies included both measures of upper and lower body strength, or measured muscle strength in different positions. Thirty-eight g-indexes were computed from the 11 investigations. Effect sizes were coded such that positive numbers reflected improvements in performance, and negative numbers reflected deterioration in performance. For each dependent measure, the Hedges g_i and the 95% confidence interval (CI) were reported. Effect size indexes of 0.20 to 0.49 were

considered small, 0.50 to 0.79 medium, and ≥ 0.80 large, based on criteria developed by Cohen.¹⁷

Several measures of sensitivity were completed. Heterogeneity among g-indexes was assessed visually using Galbraith²⁰ plots at the H_T statistics.¹⁸ As noted above, multiple g-indexes were contained in the 11 trials. To examine bias due to lack of independent data points, all g-indexes within a particular trial were averaged. This produced one g-index per investigation ($N = 11$). These g-indexes were compared to the 38 g-indexes generated across the 11 studies. We used the Shapiro-Wilks test of normality to assess whether including multiple g-indexes from a single study distorted the expected normality of effects. We also investigated the effect of any single study on the results by sequentially removing studies, one at a time, and reanalyzing the results. We tested for publication bias using the test for funnel plot asymmetry²¹ and explored other potential sources of bias or heterogeneity by examining publication date, length of study and design quality ratings.

RESULTS

The search yielded a total of 65 non-overlapping reports that were broadly construed as potentially relevant to examining the effectiveness of androgen treatment to improve physical performance in older men. The abstracts of all studies were initially examined by one author (MO) to determine if they met basic inclusion criteria. Forty-four studies were eliminated after review of the abstract. The full reports for the remaining 21 studies were examined by two raters (MO, AO) to determine the appropriateness of the study for further analysis. An additional 10 studies were eliminated following examination of the complete articles, leaving 11 studies for analysis. The study selection process is presented in Figure 1.

A total of 474 older men participated in the 11 trials included in the analysis. The mean age of participants was 69.1 years ($SD = 3.3$). The year of publication ranged from 1992 to 2005. Descriptive information for each study is included in Table 1. Ten studies provided baseline and post-treatment levels of serum testosterone (see Table 1). It was not possible to accurately compare changes in serum testosterone levels from baseline to post-treatment across studies due to the wide variation in length of treatment. There were also differences in how serum testosterone levels were determined. Some studies reported averages, while others reported lowest recorded level.

Individual g-indexes were adjusted for sample size using the modification of the Mantel-Haenzel method.^{19,22-32} The mean unadjusted g-index for the 11 studies (one effect size per study) was 0.58, (95% CI, 0.22-0.93) compared with a mean of 0.53 (95% CI, 0.21-0.86) when the 38 individual effect sizes were considered as the unit of analysis. The combined results of this analysis suggest that there were no substantial differences between any of the mean values, based on whether the aggregation occurred across the study or the individual effect size. The homogeneity statistic (H_T) was computed for the set of g-indexes and indicated that the amount of variability in the collection of g-indexes exceeded that which would have been expected by chance ($p < .05$). Subsequent procedures were completed using the individual g-indexes as the unit of analysis because it allowed investigation of covariation between outcome (as measured by g-indexes) and other design or study characteristics.

Table 2 contains mean g-indexes for the primary dependent variable (muscle strength) and selected study characteristics including the method of testosterone/DHT administration, whether the design was intention-to-treat, the amount of attrition, and overall design quality. The design quality variable was dichotomized into high and low, based on the design rating system described in the methods section. Studies with a total design rating score of < 10 were rated as low (range of design rating scores 0-13).

Examination of g-indexes for muscle strength indicates that the effect sizes were in the small to medium range and were larger (mean $g_i = 0.63$) for measures of lower extremity muscle strength. One investigation²² reported a measure of “total body strength” defined as the sum of 6; one repetition maximum weight-lifting exercises involving four upper body and two lower body movements ($g_i = 0.54$). There was a substantial difference in mean g-index by method of administration with application by injection producing a g-index of 0.95 (95% CI, 0.33-1.58), compared to g-indexes of 0.26 (95% CI, 0.08-0.42) and -0.21 (95% CI, -1.40-1.20), respectively, for topical and oral administration of testosterone.

The remaining variables in Table 2 examine the relationship between g-index and study design. A strong relationship was found between mean g-index and subject attrition. The g-index values ranged from a mean of 1.27 for studies with no attrition to 0.10 for subjects reporting attrition levels > 10%. Studies using an intention-to-treat design and analysis also reported substantially smaller g-indexes than studies not using an intention-to-treat design. Finally, trials rated as high quality were associated with a smaller mean g-index (0.30) than those with a lower quality design rating (0.64).

Additional information on the relationship between study quality and g-index was obtained by generating a funnel plot of g-indexes (x-axis) plotted against the g-index standard error (y-axis). Funnel plots are a visual tool for investigating publication and other bias in meta-analysis.³³ They are simple scatterplots of the treatment effects estimated from individual studies (horizontal axis) against a measure of study size or variability (vertical axis). The name “funnel plot” is based on the precision in the estimation of the underlying treatment effect increasing as the sample size and or variability of component studies increases. Therefore, in the absence of bias, results from small studies will scatter widely at the bottom of the graph, with the spread narrowing among larger studies. Examination of Figure 2 reveals that larger g-indexes are associated with greater standard error. Four studies with the largest g-indexes were among the investigations with the largest standard error values. Sensitivity analysis revealed that the overall mean g-index decreased substantially when the Ferrando et al.²⁵ study was removed. The mean g-index decreased from 0.53 to 0.23 (CI_{95%} = 0.09, 0.38) (median = 0.20). The relationship between g-indexes and design characteristics remained similar, but the magnitude of the relationship decreased, although it remained statistically significant ($p < .05$).

We examined each study for the reporting of adverse events. All 11 studies monitored subjects for adverse events. The definition for adverse events varied widely across the trials. Some studies provided operational definitions (e.g., increased PSA or cholesterol levels) and reported the number of events. Other investigations included broad statements such as worsened knee arthritis.²⁷ One investigation referred to adverse events as ‘severe’ or ‘not severe’ but did not define the adverse event.³² Elevated PSA levels or prostate disease were reported in three studies.^{26,28,30} Four investigations stated that no adverse events were observed in older men receiving testosterone/DHT therapy or placebo.^{22,24,25,31}

DISCUSSION

We examined the findings from 11 randomized-clinical trials using the methods of meta-analysis to determine if androgen treatment (testosterone/DHT) increased strength in men 65 years of age and above. To our knowledge, this is the first quantitative research synthesis focused on evaluating the impact of testosterone/DHT to increase strength in older men using only randomized, double-blind trials. Previous reviews of hormone replacement research including a recent meta-analysis in *Clinical Endocrinology*¹ have focused on multiple outcomes including osteopenia, frailty, insulin sensitivity, body composition, and sexual function, as well as muscle strength.^{8,9} These reviews included results from epidemiological

and observational studies, investigated different hormones (e.g., human growth hormone) or combinations of hormones, and contained findings from both younger and older men.

We found a moderate increase in overall muscle strength in subjects receiving testosterone/DHT therapy versus those receiving a placebo. The mean g-index of 0.53 can be converted to U_3 of 69.3.¹⁷ A U_3 value of 69.3 indicates that the average subject in the treatment condition receiving testosterone performed better than approximately 19.3% of the subjects in the placebo conditions. The mean g-index of 0.53 was in the range Cohen¹⁷ considers a medium treatment effect. The overall effect was influenced by the results from a single investigation that produced 5 of the 38 g-indexes. The investigation by Ferrando et al.²⁵ included 12 males, 64 to 71 years of age, who received weekly injections of testosterone for one month and then biweekly (adjusted doses) for five additional months. Muscle strength was measured in the upper (two measures) and lower (three measures) extremities using cybex exercise equipment. The Ferrando et al. study is the only investigation in which the testosterone was injected and the dosage was individually calibrated after a one-month period of weekly injections. The study continued for six months. One other investigation using testosterone injections (200 mg) and following subjects for six months²⁹ also produced a large g-index (> 0.77). Injections of 100 mg to 200 mg of testosterone produce suprphysiological levels of testosterone for several days following administration and may influence muscle mass and strength in older men in ways that are currently unknown.

The remaining studies used a different method of administration, shorter duration of treatment, or lower dosage of testosterone/DHT. The combination of individually adjusted dosage given over a period of six months and administered by injection appears to be the most effective protocol for increasing strength, but requires further study.

We found g-index differences between upper and lower extremity muscle strength, method of testosterone administration, and design characteristics. G-indexes were larger for lower extremity and total body strength measures than for upper extremity strength measures. Only one study reported a measure of total body strength.²² As noted above, effect size values were substantially larger for studies involving testosterone administered by injection versus topical or oral application. This relationship may have been confounded by dosage. We recorded information on dosage when reported; however, the variability in how dosage was determined, managed and reported prevented us from being able to accurately analyze dosage in relation to method of administration. This is also an area that requires further investigation.

All studies included in the meta-analysis were randomized clinical trials with blind recording and placebo comparison. Despite the high standards associated with double-blind RCTs, we found differences in g-indexes related to research design. Studies with zero attrition reported larger g-indexes than investigations with attrition, and this relationship appeared to be linear. In a related finding, studies using an intention-to-treat analysis reported a smaller mean g-index than studies not using an intention-to-treat analysis. Finally, studies rated as high in design quality (score of 10 or above out of a possible score of 13 on the design rating scale) reported g-indexes 53% lower than studies rated poorer in design quality. Numerous clinical and biomedical investigators have found that studies with lower design quality frequently report larger effect sizes than well-controlled RCTs.¹⁶ Our findings support the argument by Stock and others^{16,34} that even within RCTs, it is important to examine the potential impact of design quality variables on study outcome.

The strengths of our investigation include the focus on randomized clinical trials and the use of a well-defined outcome measure – muscle strength. Each statistical comparison involved testosterone/DHT in contrast to a placebo. All studies were double blind. The major limitations of the investigation were the small number of studies that met the inclusion criteria and lack

of data when statistically non-significant results were reported. The eleven studies included in the meta-analysis contained 38 statistical tests examining the effectiveness of testosterone/DHT therapy on muscle strength in older men. This number of tests did not provide sufficient data to conduct statistical sub-analyses of possible interactions between study outcomes, subject demographics and design characteristics. Our examination of these relationships was limited to descriptive statistics (see Table 2).

Another limitation was incomplete reporting of information in the primary studies. Several investigations failed to report results in sufficient detail to estimate a non-zero effect size. Some trials were more defective than others in not reporting important information. Other reviewers have encountered this problem of missing data or incomplete statistical information and commented on its possible effects.³⁵ Authors may be less likely to report specific data for analyses yielding non-significant results. To the extent that investigators fail to report results of non-significant outcomes, a degree of systematic bias is introduced regarding trial findings. We dealt with this problem by assigning a g-index of 0.00 to any statistically, non-significant comparison for which incomplete information was included in the primary investigation. This is a conservative correction and may have led to an underestimation of the overall mean g-index.

Despite these limitations, our investigation provides information regarding an important topic where research is needed to help inform clinical decision making and planning for future clinical trials. Previous reviews evaluating the effectiveness of testosterone/DHT therapy to increase muscle strength in older men have produced conflicting or inconclusive results. For example, in a recent clinical review paper, Liu and colleagues³⁶ note that previous studies “show androgen replacement in older men increases muscle and reduces fat mass to a small degree, but to date has not improved muscle strength.” This conclusion was based on studies with small sample sizes that produced statistically, non-significant results. Many of the individual studies had low statistical power which contributed to the lack of statistically significant findings.³⁶ One advantage of the meta-analysis approach is that it increases statistical sensitivity by synthesizing the results from multiple studies examining a similar research question. The IOM report, *Testosterone and Aging*,⁴ makes several recommendations regarding the need for clinical trials of testosterone therapy in older men. The IOM report argues that this effort begin with short-term trials to determine benefit. Muscle weakness and frailty are specifically identified as priority outcomes in conducting these trials.⁴ Our results suggest that issues such as method of administration, duration, individually adjusted dosage, and attrition are important characteristics to consider in planning future research on the efficacy of testosterone/DHT therapy in older men.

ACKNOWLEDGEMENTS

Financial Disclosure(s):

This research was supported by funds from the National Institute on Aging grant K02 AG01973 (K. J. Ottenbacher) and the National Center for Medical Rehabilitation Research grant K01 HD046682 (G. V. Ostir). There are no conflicts of interest.

Author Contributions:

Kenneth J. Ottenbacher: Responsible for concept and overall design, identification of variables, writing results and discussion section.

Margaret E. Ottenbacher: Responsible for data collection, coding, interpretation and writing introduction and discussion.

Allison J. Ottenbacher: Responsible for data collection, coding, assisting with analysis, reviewing and revising drafts.

Ana Alfaro Acha: Responsible for data interpretation and reviewing, revising drafts and adding new references and information to the discussion.

Glenn V. Ostir: Responsible for data interpretation, and reviewing and revising entire manuscript.

Sponsor's Role:

National Institute on Aging and the National Center for Medical Rehabilitation Research had no role in the design or preparation of the manuscript.

REFERENCES

1. Isidori A, Giannetta E, Fianfrilli D, et al. Effects of testosterone on sexual function in men: results of a meta analysis. *Clin Endocrinol* 2005;63:381–394.
2. Tan RS, Salazar JA. Risks of testosterone replacement therapy in ageing men. *Expert Opin Drug Saf* 2004;3:599–606. [PubMed: 15500418]
3. Vastag B. Many questions, few answers for testosterone replacement therapy. *JAMA* 2003;289:971–972. [PubMed: 12597734]
4. Liverman, C.; Blazer, D. Testosterone in Aging: Directions for Clinical Research. National Academy Press; Washington, DC: 2005.
5. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA* 2002;288:321–333. [PubMed: 12117397]
6. Wassertheil-Smoller S, Hendrix SL, Limacher M, et al. Effect of estrogen plus progestin on stroke in postmenopausal women: the Women's Health Initiative: a randomized trial. *JAMA* 2003;289:2673–2684. [PubMed: 12771114]
7. Bhasin S. Testosterone supplementation for aging-associated sarcopenia. *J Gerontol A Biol Sci Med Sci* 2003;58:1002–1008. [PubMed: 14630881]
8. Asthana S, Bhasin S, Butler RN, et al. Masculine vitality: pros and cons of testosterone in treating the andropause. *J Gerontol A Biol Sci Med Sci* 2004;59:461–465. [PubMed: 15123756]
9. Gruenewald DA, Matsumoto AM. Testosterone supplementation therapy for older men: potential benefits and risks. *J Am Geriatr Soc* 2003;51:101–15. [PubMed: 12534854]
10. Borst SE. Interventions for sarcopenia and muscle weakness in older people. *Age Ageing* 2004;33:548–555. [PubMed: 15385272]
11. Morley JE, Baumgartner RN, Roubenoff R, et al. Sarcopenia. *J Lab Clin Med* 2001;137:231–243. [PubMed: 11283518]
12. Bhasin S, Tenover JS. Age-associated sarcopenia--issues in the use of testosterone as an anabolic agent in older men. *J Clin Endocrinol Metab* 1997;82:1659–1660. [PubMed: 9177358]
13. Sullivan DH, Roberson PK, Johnson LE, et al. Effects of muscle strength training and testosterone in frail elderly males. *Med Sci Sports Sci* 2005;37:1664–1672.
14. Sackett, DL.; Straus, S.; Richardson, S., et al. Evidence-Based Medicine: How to Practice and Teach EBM. 2nd ed.. Churchill Livingstone; Edinburgh, UK: 2000.
15. Lipsey, MW.; Wilson, DB. Practical Meta-Analysis. Sage Publications; Thousand Oaks, Calif.: 2001.
16. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. *Medical. Stat Med* 1989;8(Part 1):441–454.
17. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed.. Lawrence Erlbaum; Hillsdale, NJ: 1988.
18. Hedges, LV.; Olkin, I. Statistical Methods for Meta-Analysis. Academic Press; Orlando, FL: 1985.
19. Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis* 1985;27:335–371. [PubMed: 2858114]
20. Galbraith R. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;7:889–894. [PubMed: 3413368]
21. Egger M, Smith GD, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629–634. [PubMed: 9310563]

22. Blackman MR, Sorkin JD, Munzer T, et al. Growth hormone and sex steroid administration in healthy aged women and men: a randomized controlled trial. *JAMA* 2002;288:2282–2292. [PubMed: 12425705]
23. Brill KT, Weltman AL, Gentili A, et al. Single and combined effects of growth hormone and testosterone administration on measures of body composition, physical performance, mood, sexual function, bone turnover, and muscle gene expression in healthy older men. *J Clin Endocrinol Metab* 2002;87:5649–5657. [PubMed: 12466367]
24. Clague JE, Wu FC, Horan MA. Difficulties in measuring the effect of testosterone replacement therapy on muscle function in older men. *Int J Androl* 1999;22:261–265. [PubMed: 10442299]
25. Ferrando AA, Sheffield-Moore M, Yeckel CW, et al. Testosterone administration to older men improves muscle function: molecular and physiological mechanisms. *Am J Physiol Endocrinol Metab* 2002;282:E601–E607. [PubMed: 11832363]
26. Kenny AM, Prestwood KM, Gruman CA, et al. Effects of transdermal testosterone on bone and muscle in older men with low bioavailable testosterone levels. *J Gerontol A Biol Sci Med Sci* 2001;56:M266–M272. [PubMed: 11320105]
27. Ly LP, Jimenez M, Zhuang TN, et al. A double-blind, placebo-controlled, randomized clinical trial of transdermal dihydrotestosterone gel on muscular strength, mobility, and quality of life in older men with partial androgen deficiency. *J Clin Endocrinol Metab* 2001;86:4078–4088. [PubMed: 11549629]
28. Page ST, Amory JK, Bowman FD, et al. Exogenous testosterone (T) alone or with finasteride increases physical performance, grip strength, and lean body mass in older men with low serum T. *J Clin Endocrinol Metab* 2005;90:1502–1510. [PubMed: 15572415]
29. Sih R, Morley JE, Kaiser FE, et al. Testosterone replacement in older hypogonadal men: a 12-month randomized controlled trial. *J Clin Endocrinol Metab* 1997;82:1661–1667. [PubMed: 9177359]
30. Snyder PJ, Peachey H, Hannoush P, et al. Effect of testosterone treatment on body composition and muscle strength in men over 65 years of age. *J Clin Endocrinol Metab* 1999;84:2647–2653. [PubMed: 10443654]
31. Tenover JS. Effects of testosterone supplementation in the aging male. *J Clin Endocrinol Metab* 1992;75:1092–1098. [PubMed: 1400877]
32. Wittert GA, Chapman IM, Haren MT, et al. Oral testosterone supplementation increases muscle and decreases fat mass in healthy elderly males with low-normal gonadal status. *J Gerontol A Biol Sci Med Sci* 2003;58:618–625. [PubMed: 12865477]
33. Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *J Clin Epidemiol* 2001;54:1046–1055. [PubMed: 11576817]
34. Stock, WA. Systematic coding for research synthesis. In: Cooper, HM.; Hedges, LV., editors. *The Handbook of Research Synthesis*. Russell Sage Foundation; New York: 1994. p. 125-138.
35. Cooper, HM. *Synthesizing Research: A Guide for Literature Reviews*. Sage Publications; Thousand Oaks, Calif.: 1998.
36. Liu PY, Swerdloff RS, Veldhuis JD. Clinical review 171: The rationale, efficacy and safety of androgen therapy in older men: future research and current practice recommendations. *J Clin Endocrinol Metab* 2004;89:4789–4796. [PubMed: 15472164]

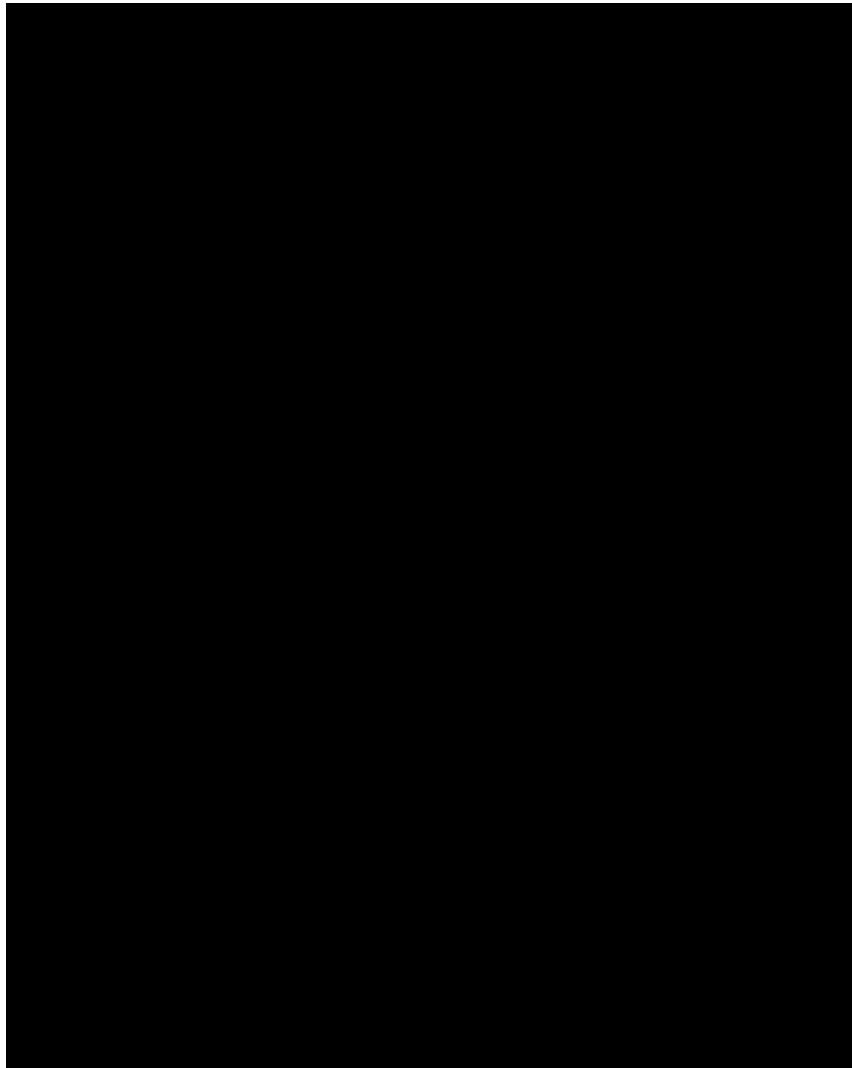


Figure 1.
Process of study selection.

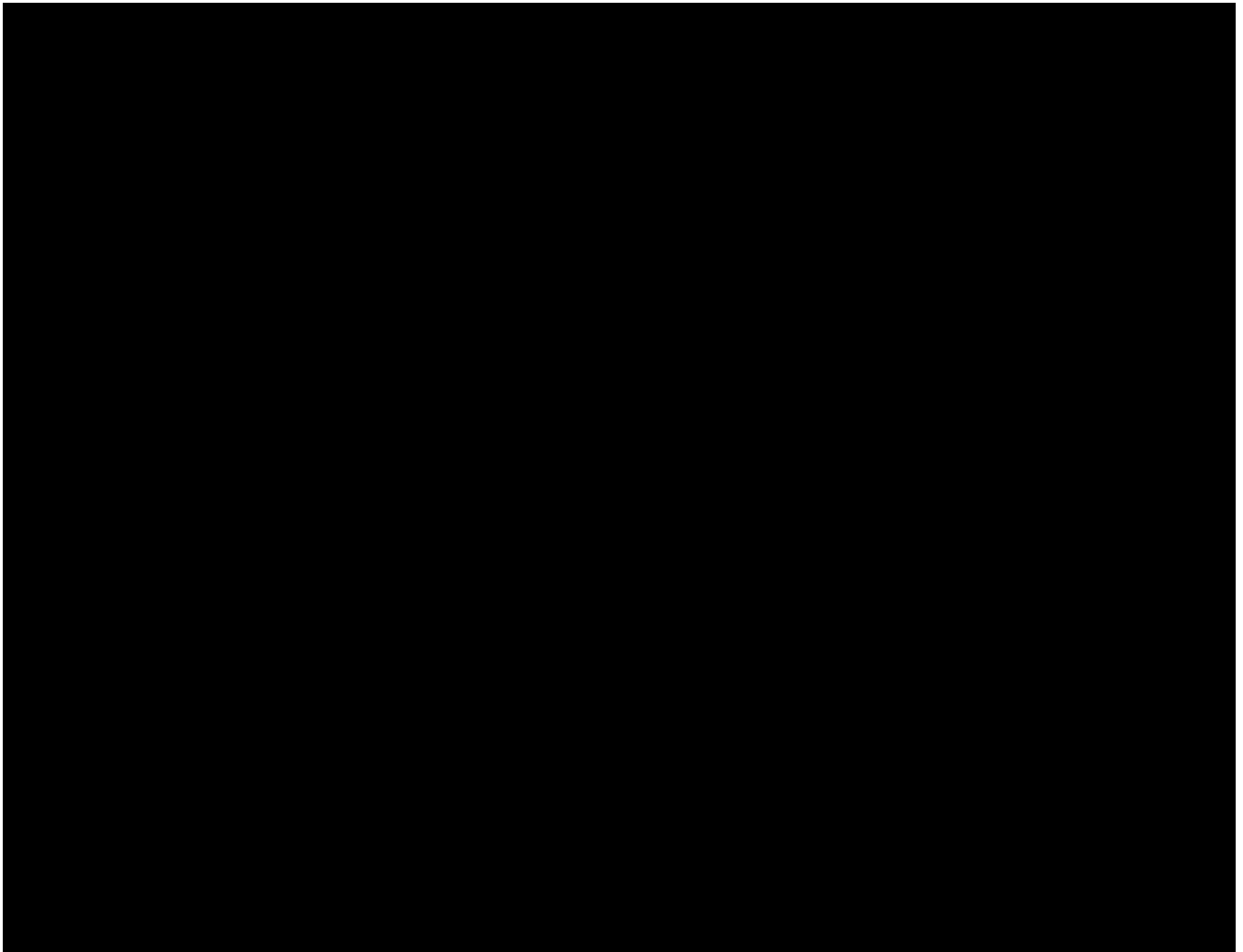


Figure 2. Funnel plot examining relationship between effect size (g-index) and standard error in 38 effect size values comparing androgen treatment to placebo in men > 65 years of age. Plot shows that studies with largest standard errors tend to be associated with larger effect sizes (g-indexes).

Table 1

Listing of randomized trials included in the meta-analysis of androgen treatment for muscle strength in older men.

Reference	N	Intervention & Baseline T ng/ml	Age (Mean Yrs.)	Design Quality*	Duration	Type of Strength Measure
22	38	Injection 100 mg biweekly; 5.9 ± NR	70	11	26 wks	Total body strength
23	10	Patch 5.0 mg Daily; NR	68	7	4 wks	Knee extension concentric Knee extension eccentric Knee flexion concentric Knee flexion eccentric
24	14	Injection 200 mg biweekly; 3.3 ± 0.5	67	11	12wks	Vertical step height Leg extensor Knee flexor Knee extensor Hand grip
25	12	Injection dosage adjusted; 3.9 ± 0.6	67	8	24 wks	Leg extension Leg curl Tricep extension Bicep curl Knee extension
26	67	Patch (2) 2.5 mg daily; 3.9 ± 1.7	76	9	52 wks	Leg extension
27	37	Gel 70 mg daily; 4.3 ± 0.9	68	7	12 wks	Shoulder non dominant Shoulder dominant extension Knee flexion dominant Knee extension dominant Knee flexion non dominant Knee extension non dominant
28	48	Injection 200 mg biweekly; 2.9 ± 0.5	71	9	156 wks	R hand grip L hand grip Ankle Knee Hand grip
29	32	Injection 200 mg biweekly; 2.9 ± 0.3	66	8	52 wks	Hand grip
30	108	Patch 6 mg/day daily; 3.7 ± 0.8	>65	10	156 wks	Knee extension 60 Knee extension 180 Knee flexion 60 Knee flexion 180 Hand grip
31	13	Injection 100 mg weekly; 3.4 ± 0.1	67	8	12 wks	R hand grip L hand grip
32	76	Oral 80 mg twice a day; 4.9 ± 1.3	69	9	52 wks	R hand grip L hand grip Quadricep Calf

* Score based on rating obtained from design quality form. Scores range from 0 to 13 with higher scores indicating better design quality. NR = not reported.

Table 2

Mean effect size values and descriptive statistics for outcome measures and study design characteristics.

Variable	N of Effect Sizes	Mean Effect Size	SE	95% CI
Strength				
Upper extremity	17	0.47	0.17	0.12-0.84
Lower extremity	20	0.63	0.31	0.03-1.28
Total body	1	0.54	NA	NA
Method of Admin.				
Topical	16	0.26	0.08	0.08-0.42
Injection	18	0.95	0.30	0.33-1.58
Oral	4	-0.21	0.38	-1.40-1.02
Attrition				
0	12	1.27	0.41	0.36-2.18
1-10%	11	0.33	0.12	0.07-0.59
>10%	15	0.10	0.12	-0.15-0.36
Intention to treat				
Yes	13	0.15	0.13	-0.13-0.43
No	20	0.89	0.27	0.32-1.46
Unknown	5	0.39	0.14	0.01-0.79
Design Quality *				
High	11	0.30	0.07	0.13-0.46
Low	27	0.64	0.22	0.17-1.09

* High = Overall design quality score of 10 or above; Low = overall design quality score of < 10. Range of scores 0 to 13.