A Meta-Analysis of the Diagnostic Accuracy of Hounsfield Units on Computed Topography Relative to Dual-Energy X-ray Absorptiometry for the Diagnosis of Osteoporosis in the Spine Surgery Population

Daniel P Ahern[1,2,3], Jake McDonnell[4], Mathieu Riffault[3,5,6], Shane Evans[7], Scott C Wagner[8], Alexander R Vaccaro[9], David A Hoey[3,5,6], Joseph S Butler[2,7]

1. School of Medicine, Trinity College Dublin, Dublin, Ireland
2. National Spinal Injuries Unit, Department of Trauma & Orthopaedic Surgery, Mater Misericordiae University Hospital, Dublin, Ireland
3. Trinity Centre for Biomedical Engineering, Trinity Biomedical Sciences Institute, Trinity College Dublin, Dublin 2 D02 R590, Ireland
4. Royal College of Surgeons in Ireland, St. Stephen's Green, Dublin, Ireland
5. Department of Mechanical and Manufacturing Engineering, School of Engineering, Trinity College Dublin, Dublin 2 D02 DK07, Ireland
6. Advanced Materials and Bioengineering Research Centre, Trinity College Dublin & RCSI, Dublin 2 D02 VN51, Ireland
7. School of Medicine and Medical Science, University College Dublin, Dublin, Ireland
8. Department of Orthopaedic Surgery, Walter Reed National Military Medical Center, Bethesda, Maryland, USA
9. Department of Orthopedic Surgery, Rothman Institute, Thomas Jefferson University, Philadelphia, USA.

Conflict of Interest: The authors declare no conflict of interest.

Funding Disclosures: Dr Daniel Ahern is in receipt of a PhD stipend from the Irish Research Council (Government of Ireland Postgraduate Scholarship: GOIPG/2019/4469). No other study specific funding disclosures or biases to declare.

Corresponding author:
Dr. Daniel P Ahern
National Spinal Injuries Unit,
Mater Misericordiae University Hospital,
Eccles Street,
Dublin 7, Ireland.

D07 AX57

Email: daniel.ahern@umail.ucc.ie

Phone: 00 353 86 8617463

Abstract

Background

The preoperative identification of osteoporosis in the spine surgery population is of crucial importance. Limitations associated with dual-energy x-ray absorptiometry(DXA), such as access and reliability, have prompted the search for alternative methods to diagnose osteoporosis. The Hounsfield Unit (HU), a readily available measure on computed tomography(CT), has garnered considerable attention in recent years as a potential diagnostic tool for reduced bone mineral density (BMD). However, the optimal threshold settings for diagnosing osteoporosis have yet to be determined.

Methods

We selected studies that included comparison of the HU (index test) with DXA evaluation (reference test). Data quality was assessed using the standardised QUADAS-2 criteria. Studies were characterised into 3 categories, based on the threshold of the index test used with the goal of obtaining a high sensitivity, high specificity or balanced sensitivity-specificity test.

Results

9 studies were eligible for meta-analysis. In the high specificity group, the pooled sensitivity was 0.652 (95% CI 0.526 – 0.760), specificity 0.795 (95% CI 0.711 – 0.859) and diagnostic odds ratio was 6.652 (95% CI 4.367 – 10.133). In the high sensitivity group, the overall pooled sensitivity was 0.912 (95% CI 0.718 – 0.977), specificity was 0.67 (0.57 – 0.75) and diagnostic odds ratio was 19.424 (5.446 – 69.275). In the balanced sensitivity-specificity group, the overall pooled sensitivity was 0.625 (95% CI 0.504 – 0.732), specificity was 0.914 (0.823 – 0.960) and diagnostic odds ratio was 14.880 (7.521 – 29.440). Considerable heterogeneity existed throughout the analysis.

Conclusion

In conclusion, the HU is a clinically useful tool to aide in the diagnosis of osteoporosis. However, the heterogeneity seen in this study warrants caution in the interpretation of results. We have demonstrated the impact of differing HU threshold values on the diagnostic ability of this test. We would propose a threshold of 135 HU to diagnose OP. Future work would investigate the optimal HU cut-off to differentiate normal from low BMD.

Introduction

Osteoporosis (OP) is a systemic skeletal disease characterised by low bone mass and a progressive microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increased risk of fracture[1]. Osteoporotic fractures result in a considerable socioeconomic and healthcare burden, due to the associated loss of independence and increased mortality rate[2-6]. Furthermore, they are associated with a significant complication profile. Notably, the risk of pseudarthrosis and subsequent mechanical failure has been reported to be as high as 35 % in the setting of osteoporosis, and this is of particular concern in an elderly patient cohort[3-5]. Moreover, the decreased pull-out strength, cut-out and insertional torque, associated with instrumenting osteoporotic patients results in an increased risk of perioperative vertebral fracture and postoperative instrumentation failure[5]. Therefore, the preoperative diagnosis of OP is crucial to ensure timely preoperative optimisation and aide in surgical planning[5,6].

The gold standard method to diagnose OP, as defined by the World Health Organisation, is a T-score of - 2.5 (i.e. more than 2.5 standard deviations below the average of a 25 year old adult) obtained by dual-energy X-ray absorption densitometry (DXA)[7]. However, access to DXA varies considerably internationally[8] and potential sources of error exist through improper patient positioning and scan interpretation[9]. These limitations have prompted the search for other techniques to diagnose OP. The Hounsfield Unit (HU), described by Schreiber et al[10] has recently emerged as a readily available alternative measure of bone mineral density on computed tomography (CT). In recent years, there have been a number of clinical studies exploring the diagnostic utility of HU. A considerable portion of the spine surgery population undergo CT imaging as routine preoperative planning to accurately assess surgical anatomy for pedicle screw placement, providing an ideal opportunity for OP screening[11]. Moreover, in the broader healthcare context, patients commonly undergo CT imaging for a wide spectrum of clinical indications, therefore allowing for opportunistic screening.

As with all diagnostic tests, sensitivity and specificity are inextricably linked and are dependent on the set threshold of the diagnostic test. At present, there is a lack of consensus on the threshold for the HU definition of OP, with varying thresholds used throughout the literature. The purpose of this meta-analysis is to synthesise the literature to-date to ascertain the overall sensitivity, specificity, and diagnostic power of the HU for OP.

Materials and Methods

Eligibility Criteria

We assessed articles based on the following eligibility criteria. The inclusion criteria included: 1) studies involving human patients/subjects 2) comparison of lumbar spine HU measurement to DXA scores 3) for the diagnosis of osteoporosis or differentiating normal from low BMD. Exclusion criteria were as follows: 1) articles not in the English language 2) studies that employed HU measurements in spinal regions other than the lumbar spine (regarding papers that include sacral levels, only data regarding lumbar levels are included) 3) studies that calculated HU measurements using imaging modalities other than conventional CT (e.g. quantitative (Q)CT) (
Figure **1**).

Literature Search

We performed a comprehensive search for eligible articles using the Medline/PubMed database and Cochrane Collaboration to include studies up to and including 25th March 2020. Search terms included “osteoporosis”, “Hounsfield unit”, “spine”, “computed tomography”. Furthermore, the bibliographies of retrieved, full-text articles were screened and the “related articles" feature in PubMed was used to identify further eligible articles. Two authors performed the literature search.

Study Selection

Two reviewers screened titles and abstracts identified in the initial search and performed the full text review of the identified studies.

Methodological Quality Assessment

Data quality was assessed using a standardised procedure as set out in the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) criteria[12]. The QUADAS-2 is a tool for the systematic review of diagnostic accuracy studies and is comprised of four domains: patient selection, index test, reference standard, and flow and timing. In brief, patient selection refers to the methods of patient recruitment, the index and reference test domains relate to how these tests were conducted and interpreted, and finally, flow and timing assesses the interval between reference and index tests.

Statistical Analysis and Data Synthesis

Statistical analysis was performed according to the Cochrane guidelines for diagnostic test accuracy (DTA) reviews[13]. The purpose of diagnostic investigations is to ascertain whether or not an individual has a particular condition or disease. Statistically, the most common measures used to gauge the accuracy of these tests are sensitivity (the proportion of those with the disease who have an abnormal result) and specificity (the proportion of those without the disease who have a normal result). Diagnostic accuracy studies compare the investigation of interest (index test) to a gold standard (reference) test. Therefore, a DTA review aims to synthesize the results of individual studies comparing the same diagnostic tests, in order to make sense of conflicting results, ultimately to better inform clinical decision making. Here, the index test is the number of Hounsfield units (HU) measured on a region of interest encompassing the lumbar vertebral body on computed tomography, with DXA as the gold standard reference test. However, the HU is a numerical value with varying thresholds used throughout the literature to categorise patients as normal or low BMD / osteoporotic. In order for accurate comparison between studies, the meta-analysis of DTA requires a common threshold to be used in all index tests. Therefore, we grouped studies in the following categories based off common thresholds used throughout the literature: 1) studies seeking a high specificity (where the HU cut-off was set at 110 HU) 2) studies seeking a high sensitivity (where the HU cut-off was set at 150 HU) and 3) where studies sought a more balanced sensitivity-specificity approach (where the HU cut-off was set at 135 HU). Where studies used marginally different cut-off values, studies were grouped into the most appropriate category. If studies incorporated more than one cut-off, data was extracted for each appropriate category and examined individually. One study[14] separated data by gender, with differing sensitivity and sensitivity values and was therefore, treated as two individual studies for the purposes of analysis.

Statistical analysis was performed using R (version 3.6.3) as described by Shim et al[15]. In brief, the R packages “metaprop” and “metabin” were used for univariate analysis to determine sensitivity, specificity, and diagnostic odds ratio (DOR). A random effects model was chosen for the analysis. In meta-analysis, there are two statistical models to choose from: a fixed effect and a random effects model.[16] A fixed effect model is used when it is assumed that there is one true effect size and that all differences in observed effects are due to sampling error. This is appropriate when comparing studies performed on the same population. The random effects model, in contrast, is used when the true effect size may differ between studies. For example, this would be most appropriate when analysing studies involving different populations, with different comorbidities, at different institutions i.e. there is an assumption of inherent heterogeneity. While the difference in effect size may be small, if there is any difference, then a random effects model is most appropriate, and it is this model which is most commonly used in meta-analysis throughout the medical literature. The DOR, a single predictor of overall test performance, can be applied to express the strength of the association between test result and disease and is a common measure used in meta-analysis[17]. It is the ratio of the odds of testing positive in those with the disease to the odds of positivity in those without the disease. The value of the DOR can range from 0 to infinity, with a higher value indicating a better performance. A value of 1 means the test is unable to differentiate between those with and without the disease. For example, if computed tomography (CT) and magnetic resonated imaging (MRI) had DORs of 7 and 12, respectively, when seeking to diagnose a particular condition, the odds of being test positive in those with the disease would be 7 and 12 times higher, respectively, than the odds of being test positive in those without the disease. Therefore, MRI would be the superior investigation of choice. Bivariate analysis was performed using the "reitsma" function of "mada" to generate summary receiving operating characteristic (sROC) curves. Coupled forest plots were used to display the sensitivity and specificity, with their 95% confidence intervals (CI), for all included studies. Forest plots were generated using GraphPad Prism software (version 8.4.1). Visual inspection of individual forest plots was used to identify heterogeneity and bias. The degree of heterogeneity was assessed using Higgins's I^2 test. The Higgins's I^2 test quantifies the total variation across studies that is due to heterogeneity rather than chance[18]. The values of I^2 range between 0% and 100%, with 0% indicating no observed heterogeneity and greater values indicating increasing heterogeneity. While thresholds for the interpretation of the I^2 are debated and should be interpreted in the context of the meta-analysis as a whole, a suggested threshold for their interpretation is 25%, 50% and 75% as low, moderate and high heterogeneity[19]. Heterogeneity is typically lower in meta-analyses comparing discrete interventions (e.g. pharmacological intervention versus placebo/ control), whereas heterogeneity is to be expected in a DTA review, for example, Nieves Plana et al[20] demonstrated 52% of all DTA Cochrane reviews reported "moderate or extreme" heterogeneity. Summary receiver operating characteristic (SROC) curves were used to display the results of individual studies in a ROC space, each study being plotted as a single sensitivity-false positive rate point. The SROC curve is the recommended method to represent the performance of a diagnostic test from the data of a meta-analysis.[21] The traditional receiver operating characteristic (ROC) curve displays the performance of a diagnostic test by indicating the relationship between true positive rate (TPR) and false positive rate (FPR) of the test across varying thresholds. Varying the threshold will result in changes to the TPR and FPR with the "best cut-off value" as the optimal threshold with the highest TPR and lowest FPR. In the SROC curve of a meta-analysis, each point represents a separate study, and each study contributes an estimate of TPR and FPR. The SROC curve, therefore, intends to represent the relationship between TPR and FPR across the studies being analysed with the "summary estimate" point as the pooled approximation of the overall test performance.[21]

Results

Literature Search

The search results are outlined in

Figure **1**. Out of 18 potential articles screened for full-text review, 9 were eligible for meta-analysis[14,22-29]. Further details of the included studies are outlined in Table 1. Study sizes ranged from 50 to 1,867 patients with mean ages ranging from 57.6 to 72.28 years. 4 studies included spinal surgical patient cohorts and the remaining 5 studies included patients undergoing CT abdomen and pelvis for other indications. As indicated in Table 1, 4 studies analysed one lumbar level, while 5 studies evaluated multiple lumbar levels. Of those studies that assessed multiple levels[22-25,27,28], one study[27] reported specific values for the entire lumbar region based on further analysis on the lumbar level HU measurement that shared the highest correlation coefficient with respective DXA values.

Methodological Assessment

Overall, there was a low risk of bias considered in the methodological quality assessment, as demonstrated in the QUADAS-2 summary (Figure 2).

Studies with a threshold of <110 HU (High Specificity)

6 studies incorporated a threshold designed for high specificity for the differentiation of osteoporosis to non-osteoporosis. The pooled sensitivity was 0.652 (95% CI 0.526 – 0.760), specificity 0.795 (95% CI 0.711 – 0.859) and diagnostic odds ratio was 6.652 (95% CI 4.367 – 10.133). Visual inspection of the forest plots indicates significant study heterogeneity, which is confirmed with an I^2 value of 85%, 81% and 51% for sensitivity, specificity and diagnostic odds ratio, respectively (Figure 3). The summary receiver operating characteristic (SROC) curve, with 95% confidence region is illustrated in Figure 4, demonstrating an area-under-the-curve (AUC) of 0.787.

Studies with a threshold >150 HU (High Sensitivity)

5 studies used a threshold of 150 HU to differentiate between normal bone mineral density (BMD) and low BMD (osteoporotic and osteopenic). The overall pooled sensitivity was 0.912 (95% CI 0.718 – 0.977), specificity was 0.67 (0.57 – 0.75) and diagnostic odds ratio was 19.424 (5.446 – 69.275). Heterogeneity is appreciated through visual inspection of the

forest plots, in particular for sensitivity (I^2 = 96%) and diagnostic odds ratio (I^2 = 83%). The specificity forest plot had less apparent heterogeneity, which was confirmed with an I^2 value of 26% (Figure 5). The SROC curve, illustrated in Figure 6, had an AUC of 0.744.

Studies with a threshold of 135 HU (balanced sensitivity- specificity)

5 studies incorporated a threshold for a more balanced sensitivity-specificity approach for the diagnosis of osteoporosis. The overall pooled sensitivity was 0.625 (95% CI 0.504 – 0.732), specificity was 0.914 (0.823 – 0.960) and diagnostic odds ratio was 14.880 (7.521 – 29.440). Heterogeneity is appreciable on the forest plots and confirmed with I^2 values of 84%, 84% and 63% for sensitivity, specificity and DOR, respectively (Figure 7). The SROC curve, with 95% confidence region is illustrated in Figure 8, demonstrating an area-under-the-curve (AUC) of 0.831.

Discussion

In this systematic review, the diagnostic utility of HU for the diagnosis and screening of OP was examined. By employing statistical methods designed specifically for diagnostic meta-analysis, the available evidence surrounding the use of the HU was quantitatively synthesised. Overall, we found that the HU provides clinically useful information regarding BMD determination (Table 2). For accurate diagnostic test meta-analysis, the threshold between studies must be uniform. Unfortunately, it is not possible to determine pooled thresholds to maximise sensitivity and specificity. Therefore, we divided the evidence into three categories, based on the diagnostic goal of the set threshold; high specificity, high sensitivity, and balanced sensitivity/ specificity. From the methodological assessment, there was, overall, a low risk of bias appreciated by the reviewers. The concerns for potential bias included potential selection bias based on the presence of a recent DXA examination as an inclusion criterion and where the HU measurement was taken from contrast-enhanced imaging examinations, such as a CT abdomen and pelvis. It is unclear what effect, if any, this bias may have on the results of this study. There were no concerns of bias in regard to the reference DXA test or the flow and timing of the tests.

The group with a threshold cut-off of 110 HU, designed for a higher specificity test, had a pooled sensitivity of 0.652, specificity of 0.795 and DOR of 6.562. Interestingly, the group with a threshold setting of 135 HU had a higher pooled specificity value of 0.914, and DOR of 14.880 with a comparable sensitivity value of 0.625. While significant heterogeneity exists within both study groups, both thresholds provide a clinically useful specificity. A threshold cut-off of 150 HU yielded a pooled sensitivity of 0.912, superior to the thresholds of 110 HU (0.652) and 135HU (0.625). Therefore, one might consider a threshold of 150 HU a clinically effective screening test for osteoporosis.

Considerable work has been published since Schreiber et al[10] described the use of the HU as a diagnostic tool for osteoporosis in 2011. However, this is the first meta-analysis on this topic to date. One of the limitations prohibiting the advancement of this technique is the heterogeneity of thresholds used throughout the literature. This heterogeneity is partly due to the differing diagnostic utility of the test (screening and diagnosing). A systematic review by Zaidi et al[30] in 2018 proposed similar cut-offs to the thresholds used in this analysis: 110 HU for high specificity, 160 HU for high sensitivity and 135 HU for balanced sensitivity-specificity. Wagner et al[31] utilised this technique to determine that a considerable number of patients undergoing transforaminal lumbar interbody fusion (TLIF) suffered from undiagnosed osteoporosis and low BMD. A cut-off of 150HU was used to determine normal from low BMD, and a cut-off below 112.4 HU was used to diagnose frank OP versus osteopenia. In this analysis, we determined a cut-off of 150 HU as the most appropriate for a high sensitivity diagnostic approach.

The principle advantage of using the HU for the orthopaedic surgeon is the timely preoperative identification of a patient suffering from low BMD. It is a quick, easy to use and readily available technique with numerous studies reporting excellent inter-observer variability[10,14,25,32,33]. The preoperative identification of osteoporosis facilitates the medical optimisation of patients perioperatively, such as calcium and vitamin D supplementation and treatment with bisphosphonates or recombinant parathyroid hormone (rPTH)[34]. Surgical strategies employed when instrumenting an osteoporotic spine include use of longer constructs and avoiding starting or ending a construct at junctional levels to prevent segmental or junctional failure. One might consider at least 3 fixation points above and below any spinal deformity. Furthermore, when using a long fusion construct, one might consider fixation to the pelvis to maximize construct stability. Anterior column support should be considered to increase load-sharing and minimise the strain on a construct. Pedicle screws can be undertapped to increase insertional torque and pull-out strength. There may also be a role for hydroxyapatite (HA)-coated screws or fenestrated/cement augmented screws to minimise the risk of mechanical failure[35].

Another consideration when using this technique relates to the vertebra of interest measured. Mean HU values differ significantly throughout the skeleton, such as the ulna[36], carpal bones[37], femoral head[33], femoral neck[33], talus[33] and facial bones[38]. In the seminal work by Schreiber, mean HU values between the L1-L4 vertebrae were not significantly different[10]. Pickhardt et al[26], in the largest series to date on this subject, identified a trend towards differing mean attenuation values between various lumbar vertebrae, with the lowest attenuation at L3. However, this difference between vertebrae was not significant. Berger-Groch et al. demonstrated differing HU values to differentiate between normal and low BMD in the L4, L5 and S1 vertebrae[22]. Therefore, we recommend HU measurements to be taken between the L1-4 vertebrae.

Moreover, utilising HU in special populations, such as ankylosing spondylitis (AS), degenerative disease, and scoliosis may provide more accurate measurements of BMD than DXA. AS is a common skeletal inflammatory condition frequently affecting the spine, resulting in both new bone formation and an overall reduction in bone quality[39], which can impair the reliability of DXA[40]. Artefactual elevations in calcium content are seen in certain inflammatory conditions of the spine. This does not represent true BMD and can cause falsely elevated readings on DXA which may ultimately go unappreciated by the operator, leading to inaccurate interpretation of certain patients' true BMD[40]. Emohare et al[39] demonstrated the utilisation of this technique in a series of 17 patients with AS to good effect. Pappou et al[41] described the positive correlation between falsely elevated BMD scores on lumbar DXA and increasing curve angle in degenerative lumbar scoliosis.

This study has a number of limitations. Firstly, there is a relatively small number of included studies, especially when divided for subgroup analysis, that are predominantly retrospective in nature. Due to this limitation, a sensitivity and heterogeneity analysis was not performed. However, visual examination of the generated forest plots can determine which studies contributed to increased heterogeneity. Secondly, as described above, some studies used marginally different threshold values for each subgroup, we grouped studies into their most appropriate subgroup with the assumption that these differences result in negligible changes in BMD. 5 of the included studies used data generated from CT abdomen and pelvis obtained for other indications, and likely used intravenous contrast. It is not currently known what effect, if any, intravenous contrast has on spinal HU measurement. Thirdly, further analysis regarding variability in HU measurements due to different tube voltages of specific imaging systems was not possible as it was not consistently listed throughout studies. This is worth noting for further studies as variation in tube voltages has been shown to impact HU measurements[42]. There were no funding disclosures or biases which would have influenced the conception, design or interpretation of the study presented herein.

In conclusion, the HU is a clinically useful tool to aide in the diagnosis of osteoporosis. However, the heterogeneity seen in this study warrants caution in the interpretation of results. We have demonstrated the impact of differing HU threshold values on the diagnostic ability of this test. We would propose a threshold of 135 HU to diagnose OP. Future work would investigate the optimal HU cut-off to differentiate normal from low BMD.

References:

1. Christiansen C. Consensus development conference: prophylaxis and treatment of osteoporosis. *Am J Med.* 1991;90:107-110.
2. Osteoporosis prevention d, therapy. NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy. *JAMA.* 2001;285(6):785-795.
3. Ohtori S, Inoue G, Orita S, et al. Teriparatide accelerates lumbar posterolateral fusion in women with postmenopausal osteoporosis: prospective study. *Spine.* 2012;37(23):E1464-E1468.
4. Berjano P, Langella F, Damilano M, et al. Fusion rate following extreme lateral lumbar interbody fusion. *European Spine Journal.* 2015;24(3):369-371.
5. Karikari IO, Metz LN. Preventing Pseudoarthrosis and Proximal Junctional Kyphosis: How to Deal with the Osteoporotic Spine. *Neurosurgery Clinics.* 2018;29(3):365-374.
6. Park SB, Chung CK. Strategies of spinal fusion on osteoporotic spine. *Journal of Korean Neurosurgical Society.* 2011;49(6):317.
7. Organization WH. Prevention and management of osteoporosis. *World Health Organ Tech Rep Ser.* 2003;921:1-164.
8. Kanis J, Johnell O. Requirements for DXA for the management of osteoporosis in Europe. *Osteoporosis international.* 2005;16(3):229-238.
9. Watts NB. Fundamentals and pitfalls of bone densitometry using dual-energy X-ray absorptiometry (DXA). *Osteoporosis international.* 2004;15(11):847-854.
10. Schreiber JJ, Anderson PA, Rosas HG, Buchholz AL, Au AG. Hounsfield units for assessing bone mineral density and strength: a tool for osteoporosis management. *JBJS.* 2011;93(11):1057-1063.
11. Wi W, Park S-M, Shin B-S. Computed Tomography-Based Preoperative Simulation System for Pedicle Screw Fixation in Spinal Surgery. *Journal of Korean Medical Science.* 2020;35(18).
12. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine.* 2011;155(8):529-536.

13. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Cochrane handbook for systematic reviews of diagnostic test accuracy. *Version 09 0 London: The Cochrane Collaboration.* 2010.

14. Kim Y, Kim JH, Yoon SH, et al. Vertebral bone attenuation on low-dose chest CT: quantitative volumetric analysis for bone fragility assessment. *Osteoporosis International.* 2017;28(1):329-338.

15. Shim SR, Kim S-J, Lee J. Diagnostic test accuracy: application and practice using R software. *Epidemiology and health.* 2019;41.

16. Sotiriadis A, Papatheodorou S, Martins W. Synthesizing evidence from diagnostic accuracy tests: the SEDATE guideline. *Ultrasound in Obstetrics & Gynecology.* 2016;47(3):386-395.

17. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology.* 2003;56(11):1129-1135.

18. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Bmj.* 2003;327(7414):557-560.

19. Deeks JJ, Higgins JP, Altman DG, Group CSM. Analysing data and undertaking meta-analyses. *Cochrane handbook for systematic reviews of interventions.* 2019:241-284.

20. Plana MN, Pérez T, Zamora J. A meta-epidemiological study of reporting of heterogeneity measures in Diagnostic Test Accuracy (DTA) reviews. A proposal of new measures to quantify heterogeneity. *Journal of Clinical Epidemiology.* 2020.

21. Walter S. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Statistics in medicine.* 2002;21(9):1237-1256.

22. Berger-Groch J, Thiesen D, Ntalos D, Hennes F, Hartel M. Assessment of bone quality at the lumbar and sacral spine using CT scans: a retrospective feasibility study in 50 comparing CT and DXA data. *European Spine Journal.* 2020:1-7.

23. Buckens CF, Dijkhuis G, de Keizer B, Verhaar HJ, de Jong PA. Opportunistic screening for osteoporosis on routine computed tomography? An external validation study. *European radiology.* 2015;25(7):2074-2079.

24. Li Y-L, Wong K-H, Law MW-M, et al. Opportunistic screening for osteoporosis in abdominal computed tomography for Chinese population. *Archives of osteoporosis.* 2018;13(1):76.

25. Kim KJ, Kim DH, Lee JI, Choi BK, Han IH, Nam KH. Hounsfield units on lumbar computed tomography for predicting regional bone mineral density. *Open Medicine.* 2019;14(1):545-551.

26. Pickhardt PJ, Pooler BD, Lauder T, del Rio AM, Bruce RJ, Binkley N. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Annals of internal medicine.* 2013;158(8):588-595.

27. CANSU A, Atasoy D, EYÜBOĞLU İ, Karkucak M. Diagnostic efficacy of routine contrast-enhanced abdominal CT for the assessment of osteoporosis in the Turkish population. *Turkish Journal of Medical Sciences.* 2020;50(1):110-116.

28. Zou D, Li W, Deng C, Du G, Xu N. The use of CT Hounsfield unit values to identify the undiagnosed spinal osteoporosis in patients with lumbar degenerative diseases. *European Spine Journal.* 2019;28(8):1758-1766.

29. Zou D, Jiang S, Zhou S, et al. Prevalence of Osteoporosis in Patients Undergoing Lumbar Fusion for Lumbar Degenerative Diseases: A Combination of DXA and Hounsfield Units. *Spine.* 2020;45(7):E406-E410.

30. Zaidi Q, Danisa OA, Cheng W. Measurement techniques and utility of Hounsfield unit values for assessment of bone quality prior to spinal instrumentation: a review of current literature. *Spine.* 2019;44(4):E239-E244.

31. Wagner S, Kang DG, Steelman T, Helgeson MD, Lehman RA. Diagnosing the undiagnosed: osteoporosis in patients undergoing lumbar fusion. *The Spine Journal.* 2016;16(10):S301.

32. Hendrickson NR, Pickhardt PJ, del Rio AM, Rosas HG, Anderson PA. Bone mineral density T-scores derived from CT attenuation numbers (Hounsfield units): clinical utility and correlation with dual-energy X-ray absorptiometry. *The Iowa orthopaedic journal.* 2018;38:25.

33. Lee SY, Kwon S-S, Kim H, et al. Reliability and validity of lower extremity computed tomography as a screening tool for osteoporosis. *Osteoporosis International.* 2015;26(4):1387-1394.

34. Lubelski D, Choma TJ, Steinmetz MP, Harrop JS, Mroz TE. Perioperative medical management of spine surgery patients with osteoporosis. *Neurosurgery.* 2015;77(suppl_1):S92-S97.

35. Goldstein CL, Brodke DS, Choma TJ. Surgical management of spinal conditions in the elderly osteoporotic spine. *Neurosurgery.* 2015;77(suppl_1):S98-S107.

36. Wagner SC, Dworak TC, Grimm PD, Balazs GC, Tintle SM. Measurement of distal ulnar Hounsfield units accurately predicts bone mineral density of the forearm. *JBJS.* 2017;99(8):e38.

37. Johnson CC, Gausden EB, Weiland AJ, Lane JM, Schreiber JJ. Using Hounsfield units to assess osteoporotic status on wrist computed tomography scans: comparison with dual energy x-ray absorptiometry. *The Journal of hand surgery.* 2016;41(7):767-774.

38. Lee IJ, Lee JJ, Bae J-H, et al. Significance of osteoporosis in facial bone density using computed tomography. *Journal of Craniofacial Surgery.* 2013;24(2):428-431.

39. Emohare O, Cagan A, Polly Jr DW, Gertner E. Opportunistic computed tomography screening shows a high incidence of osteoporosis in ankylosing spondylitis patients with acute vertebral fractures. *Journal of Clinical Densitometry.* 2015;18(1):17-21.

40. Gregson CL, Hardcastle SA, Cooper C, Tobias JH. Friend or foe: high bone mineral density on routine bone density scanning, a review of causes and management. *Rheumatology.* 2013;52(6):968-985.

41. Pappou IP, Girardi FP, Sandhu HS, et al. Discordantly high spinal bone mineral density values in patients with adult lumbar scoliosis. *Spine.* 2006;31(14):1614-1620.

42. Garner HW, Paturzo MM, Gaudier G, Pickhardt PJ, Wessell DE. Variation in attenuation in L1 trabecular bone at different tube voltages: caution is warranted when screening for osteoporosis with the use of opportunistic CT. *American Journal of Roentgenology.* 2017;208(1):165-170.

Figure Legend

Figure 1 Flowchart of studies included for meta-analysis

Figure 2 A) QUADAS-2 summary figure and B) study specific breakdown of the QUADAS-2 evaluation tool. Study grading was based on the study reviewers’ judgement.

Figure 3 Forest plots of sensitivity, specificity and diagnostic odds ratio for studies with a 110 HU cut-off for the diagnosis of osteoporosis.

Figure 4 Summary receiver operating characteristic (SROC) curve for studies with a 110 HU cut-off for the diagnosis of osteoporosis.

Figure 5 Forest plots of sensitivity, specificity and diagnostic odds ratio for studies with a 150 HU cut-off for the diagnosis of normal versus low bone mineral density.

Figure 6 Summary receiver operating characteristic (SROC) curve for studies with a 150 HU cut-off for the diagnosis of normal versus low bone mineral density

Figure 7 Forest plots of sensitivity, specificity and diagnostic odds ratio for studies with a 135 HU cut-off for the diagnosis of osteoporosis

Figure 8 Summary receiver operating characteristic (SROC) curve for studies with a 135 HU cut-off for the diagnosis of osteoporosis

Figure 1

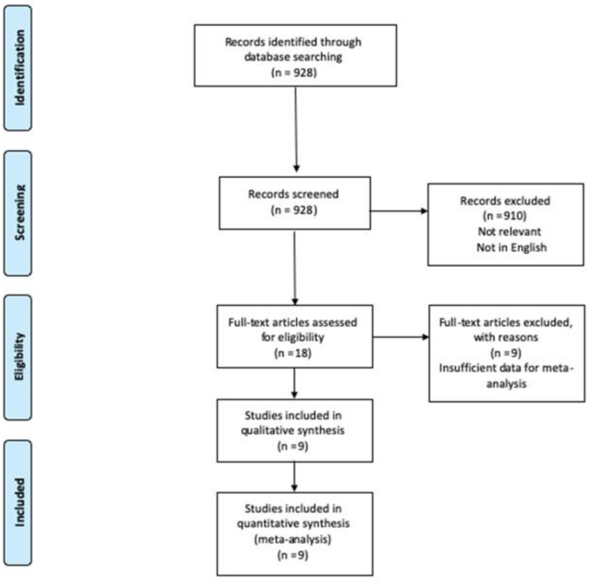
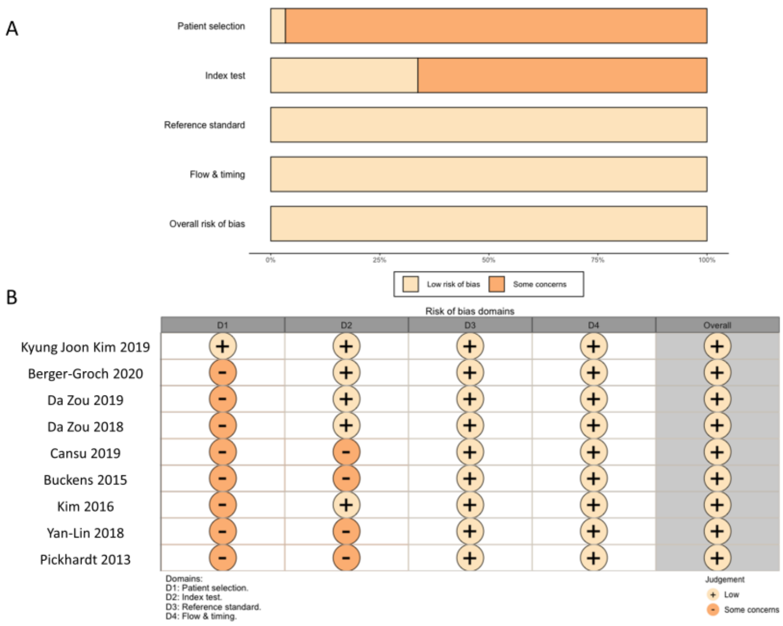

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 928)
Records screened (n = 928)
Records excluded (n =910) Not relevant Not in English
Full-text articles assessed for eligibility (n =18)
Full -text articles excluded, with reasons (n =9) Insufficient data for meta - analysis
Studies included in qualitative synthesis (n =9)
Studies included in quantitative synthesis (meta-analysis) (n =9)

Figure 2



Risk of bias domains

	D1	D2	D3	D4	Overall
Kyung Joon Kim 2019	+	+	+	+	+
Berger-Groch 2020	-	+	+	+	+
Da Zou 2019	-	+	+	+	+
Da Zou 2018	-	+	+	+	+
Cansu 2019	-	-	+	+	+
Buckens 2015	-	-	+	+	+
Kim 2016	-	+	+	+	+
Yan-Lin 2018	-	-	+	+	+
Pickhardt 2013	-	-	+	+	+

Domains:
D1: Patient selection.
D2: Index test.
D3: Reference standard.
D4: Flow & timing.

Judgement
+ Low
- Some concerns

Figure 3

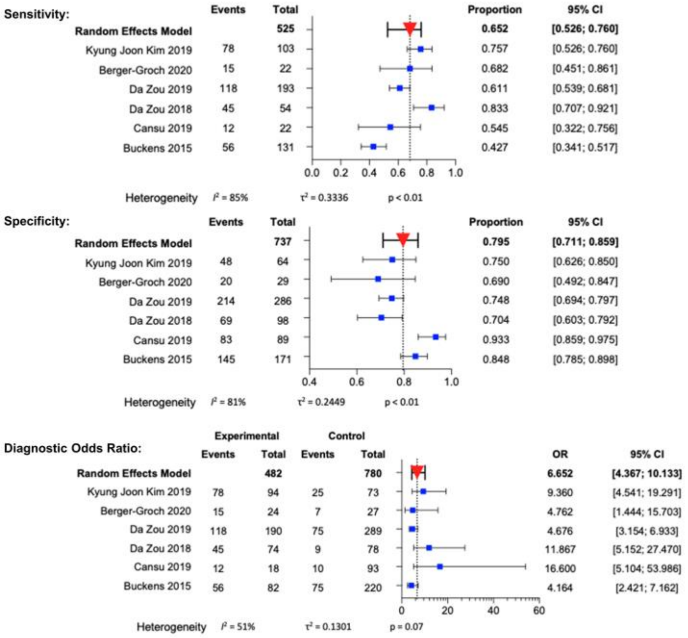

Sensitivity:
Events
Total
Proportion
95% CI
Random Effects Model
525
0.652
[0.526; 0.760]
Kyung Joon Kim 2019
78
103
0.757
[0.526; 0.760]
Berger-Groch 2020
15
22
0.682
[0.451; 0.861]
Da Zou 2019
118
193
0.611
[0.539; 0.681]
Da Zou 2018
45
54
0.833
[0.707; 0.921]
Cansu 2019
12
22
0.545
[0.322; 0.756]
Buckens 2015
56
131
0.427
[0.341; 0.517]
0.0
0.2
0.4
0.6
0.8
1.0
Heterogeneity
$I^2 = 85\%$
$\tau^2 = 0.3336$
$p < 0.01$
Specificity:
Events
Total
Proportion
95% CI
Random Effects Model
737
0.795
[0.711; 0.859]
Kyung Joon Kim 2019
48
64
0.750
[0.626; 0.850]
Berger-Groch 2020
20
29
0.690
[0.492; 0.847]
Da Zou 2019
214
286
0.748
[0.694; 0.797]
Da Zou 2018
69
98
0.704
[0.603; 0.792]
Cansu 2019
83
89
0.933
[0.859; 0.975]
Buckens 2015
145
171
0.848
[0.785; 0.898]
0.4
0.6
0.8
1.0
Heterogeneity
$I^2 = 81\%$
$\tau^2 = 0.2449$
$p < 0.01$
Diagnostic Odds Ratio:
Experimental
Control
Events
Total
Events
Total
OR
95% CI
Random Effects Model
482
780
6.652
[4.367; 10.133]
Kyung Joon Kim 2019
78
94
25
73
9.360
[4.541; 19.291]
Berger-Groch 2020
15
24
7
27
4.762
[1.444; 15.703]
Da Zou 2019
118
190
75
289
4.676
[3.154; 6.933]
Da Zou 2018
45
74
9
78
11.867
[5.152; 27.470]
Cansu 2019
12
18
10
93
16.600
[5.104; 53.986]
Buckens 2015
56
82
75
220
4.164
[2.421; 7.162]
0
20
40
60
Heterogeneity
$I^2 = 51\%$
$\tau^2 = 0.1301$
$p = 0.07$

Figure 4

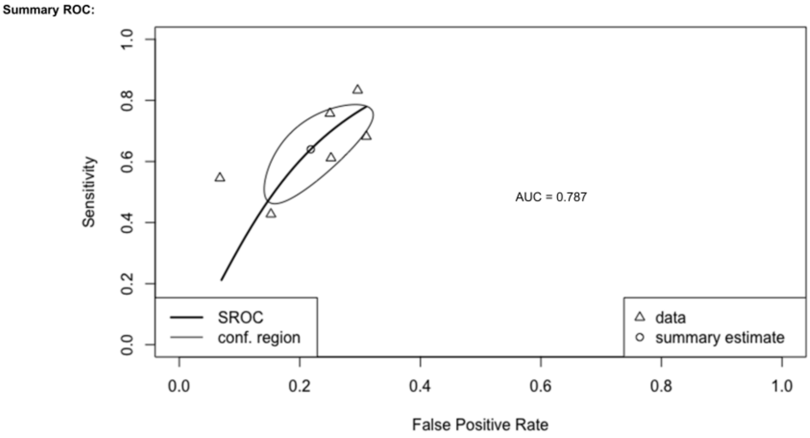

Summary ROC:
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.787
SROC
conf. region
△ data
○ summary estimate

Figure 5

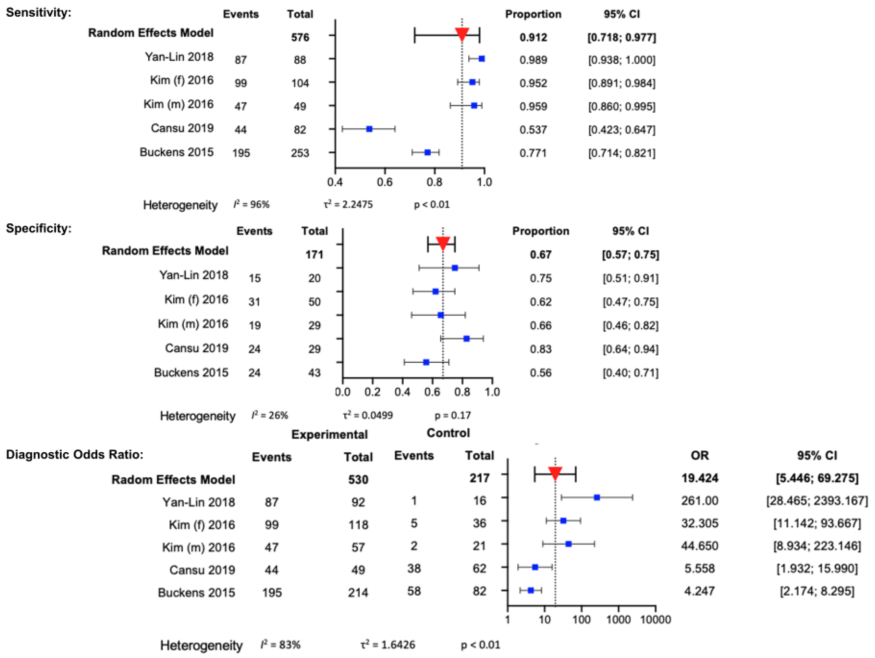

Sensitivity:
Events Total Proportion 95% CI
Random Effects Model 576 0.912 [0.718; 0.977]
Yan-Lin 2018 87 88 0.989 [0.938; 1.000]
Kim (f) 2016 99 104 0.952 [0.891; 0.984]
Kim (m) 2016 47 49 0.959 [0.860; 0.995]
Cansu 2019 44 82 0.537 [0.423; 0.647]
Buckens 2015 195 253 0.771 [0.714; 0.821]
0.4 0.6 0.8 1.0
Heterogeneity I^2 = 96% τ^2 = 2.2475 p < 0.01
Specificity:
Events Total Proportion 95% CI
Random Effects Model 171 0.67 [0.57; 0.75]
Yan-Lin 2018 15 20 0.75 [0.51; 0.91]
Kim (f) 2016 31 50 0.62 [0.47; 0.75]
Kim (m) 2016 19 29 0.66 [0.46; 0.82]
Cansu 2019 24 29 0.83 [0.64; 0.94]
Buckens 2015 24 43 0.56 [0.40; 0.71]
0.0 0.2 0.4 0.6 0.8 1.0
Heterogeneity I^2 = 26% τ^2 = 0.0499 p = 0.17
Experimental Control
Diagnostic Odds Ratio:
Events Total Events Total OR 95% CI
Radom Effects Model 530 217 19.424 [5.446; 69.275]
Yan-Lin 2018 87 92 1 16 261.00 [28.465; 2393.167]
Kim (f) 2016 99 118 5 36 32.305 [11.142; 93.667]
Kim (m) 2016 47 57 2 21 44.650 [8.934; 223.146]
Cansu 2019 44 49 38 62 5.558 [1.932; 15.990]
Buckens 2015 195 214 58 82 4.247 [2.174; 8.295]
1 10 100 1000 10000
Heterogeneity I^2 = 83% τ^2 = 1.6426 p < 0.01

Figure 6

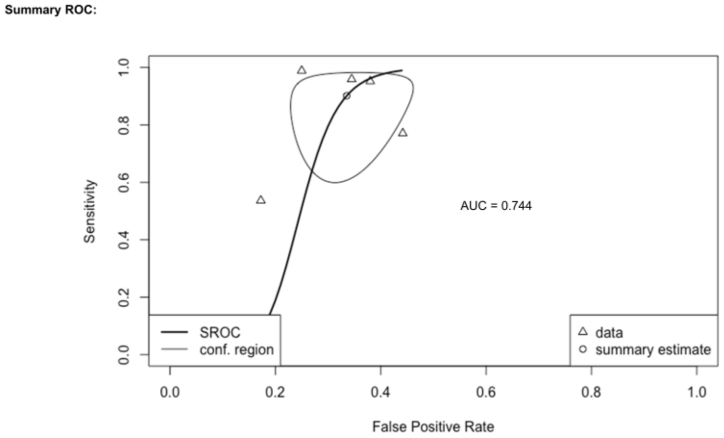

Summary ROC:
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.744
SROC
conf. region
△ data
○ summary estimate

Figure 7

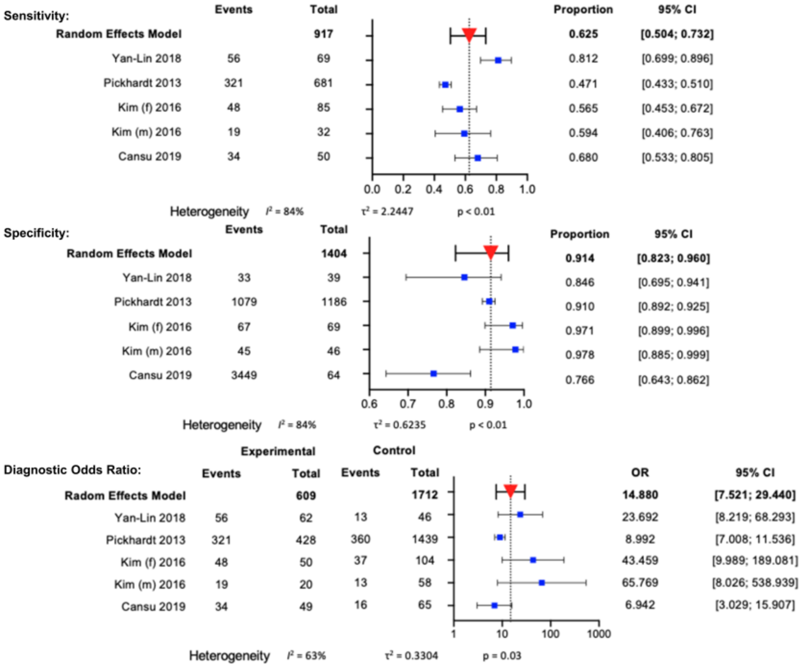

Sensitivity:
Events
Total
Proportion
95% CI
Random Effects Model 917 0.625 [0.504; 0.732]
Yan-Lin 2018 56 69 0.812 [0.699; 0.896]
Pickhardt 2013 321 681 0.471 [0.433; 0.510]
Kim (f) 2016 48 85 0.565 [0.453; 0.672]
Kim (m) 2016 19 32 0.594 [0.406; 0.763]
Cansu 2019 34 50 0.680 [0.533; 0.805]
0.0 0.2 0.4 0.6 0.8 1.0
Heterogeneity $I^2 = 84\%$ $\tau^2 = 2.2447$ $p < 0.01$
Specificity:
Events
Total
Proportion
95% CI
Random Effects Model 1404 0.914 [0.823; 0.960]
Yan-Lin 2018 33 39 0.846 [0.695; 0.941]
Pickhardt 2013 1079 1186 0.910 [0.892; 0.925]
Kim (f) 2016 67 69 0.971 [0.899; 0.996]
Kim (m) 2016 45 46 0.978 [0.885; 0.999]
Cansu 2019 3449 64 0.766 [0.643; 0.862]
0.6 0.7 0.8 0.9 1.0
Heterogeneity $I^2 = 84\%$ $\tau^2 = 0.6235$ $p < 0.01$
Experimental
Control
Diagnostic Odds Ratio:
Events
Total
Events
Total
OR
95% CI
Radom Effects Model 609 1712 14.880 [7.521; 29.440]
Yan-Lin 2018 56 62 13 46 23.692 [8.219; 68.293]
Pickhardt 2013 321 428 360 1439 8.992 [7.008; 11.536]
Kim (f) 2016 48 50 37 104 43.459 [9.989; 189.081]
Kim (m) 2016 19 20 13 58 65.769 [8.026; 538.939]
Cansu 2019 34 49 16 65 6.942 [3.029; 15.907]
1 10 100 1000
Heterogeneity $I^2 = 63\%$ $\tau^2 = 0.3304$ $p = 0.03$

Figure 8

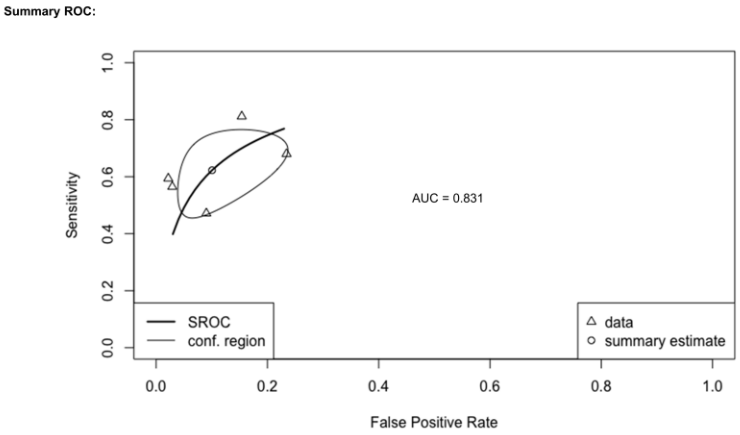

Summary ROC:
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.831
SROC
conf. region
data
summary estimate

Table 1 Characteristics of eligible studies

Study	**No. of Patients**	**Patient Population**	**Mean Age (yr.)**	**Lumbar Vertebrae Measured**	**Hounsfield Unit Cut-off**
Berger-Groch 2020[20(a)]	50	Spinal surgical patients	72.28 (range 31 – 89)	L4, L5, S1	Normal bone: L4 > 161 HU, L5 > 157 HU, S1 207 HU Osteoporotic bone: L4 < 62 HU, L5 < 58 HU, S1 < 68 HU
Buckens 2015[21 (a, b)]	302	Consecutive patients undergoing CT chest/abdomen with a recent DXA	61 (OP), 57 (Non-OP)	L1	Validation of 160 HU, 110 HU, 80 HU as thresholds
Kim 2016[18 (b, c)]	232	Consecutive patients with CT and recent DXA examination	M = 65.9 +/- 8.5, F = 64.1 +/- 9.9	L1	Male: 136.2 HU = 95% sensitivity, 77.6% specificity and 165.7 HU = 82.5% sensitivity and 90.5% specificity Female: 137.9 HU = 96% sensitivity, 64.4% specificity and 151 HU = 83.9% sensitivity and 86.1% specificity.
Yan-Lin 2018[22 (b, c)]	109	Chinese patients undergoing CT abdomen pelvis for other indications, with recent DEXA	66.8 (range 11 – 91)	L1- L5	diagnosis of OP ≤ 136 HU and exclusion ≥ 175 HU)
Kyung Joon Kim 2019[23 (a)]	331	Spinal surgery patients undergoing QCT,	NR	L1-L3	OP ≤ 95 ≤ normal

		lumbar CT and DEXA			
Pickhardt 2013[24 (c)]	1867	Patients undergoing abdominal CT and recent DXA	59.2 (+/- 12.5)	L1	135 HU 90% sensitive, 190 HU 90% specific
Cansu 2020[25 (a, b, c)]	111	Consecutive patients with a CT abdomen pelvis and recent DXA	57.6 (16 – 87)	L1-L4	High sensitivity : ≤ 170 (sensitivity 88.9%, specificity 25.3%) High specificity: ≤ 102 (sensitivity 66.7%, specificity 89.1%) Balanced: ≤121 (sensitivity 77.8%, specificity 68.9%)
Da Zou 2018[26 (a)]	152	Lumbar spinal surgery patients with recent DXA	58.5 +/- 6.4	L1 – L4	L1: 110 HU, L2 : 100 HU, L3 : 85 HU, L4: 80 HU
Da Zou 2020 [27 (a)]	479	Patients over 50 with lumbar degenerative disease	61.8 +/- 6.8	L1	L1 ≤ 110 HU

a, b, c denote studies included in a) high specificity (cut-off 110 HU) group b) high sensitivity (cut-off 150 HU) group and c) balanced cut-off (cut-off 136 HU). OP; osteoporosis.

Table 2

Table 1 Summary of Findings Table

General Information		
General Issue	What is the diagnostic performance of computed tomography (CT) derived Hounsfield Unit (HU) in diagnosing osteoporosis (OP)	
Specific Questions	What is the diagnostic performance in distinguishing OP from non-OP?	OP versus osteopenia/normal BMD
	What is the diagnostic performance in distinguishing normal BMD from low BMD?	Normal BMD versus osteopenia/OP
	What is the diagnostic performance when using a balanced threshold to distinguish OP from non-OP?	OP versus osteopenia/normal BMD
Patients	Patients with lumbar spine CT and DXA	
Index tests	CT derived Hounsfield Unit (HU)	
Reference Standard	DXA	
Quality concerns	Overall Judgement	Low risk
	Patient Selection bias	8 studies with some concerns due to retrospective nature of studies
	Index test interpretation bias	4 studies with some concerns due to use of contrast enhanced CT
	Reference test interpretation bias	Low risk
	Flow and timing selection bias	Low risk
Studies which diagnosed osteoporosis at <111 HU		
Studies	6 (1,425 patients)	
Summary results	Sensitivity 0.652 (95% CI: 0.526 – 0.76) Specificity 0.80 (95% CI: 0.71 – 0.86)	
Consequences	In a hypothetical cohort of 1,000 patients (prevalence 40%)	Correctly Classified: 741
		Underdiagnosed: 139
		Over-diagnosed: 120
Studies which diagnosed low BMD at <111 HU		
Studies	4 (754 patients)	
Summary Results	Sensitivity 0.912 (95% CI: 0.718 – 0.977) Specificity 0.67 (95% CI: 0.57 – 0.75)	
Consequences	In a hypothetical cohort of 1,000 patients (prevalence 40%)	Correctly Classified: 767
		Underdiagnosed: 35
		Over-diagnosed: 198
Studies which diagnosed osteoporosis as <135 HU		
Studies	4 (2,319 patients)	
Summary Results	Sensitivity 0.625 (95% CI: 0.504 – 0.732) Specificity 0.914 (95% CI: 0.823 – 0.960)	
Consequences	In a hypothetical cohort of 1,000 patients (prevalence 40%)	Correctly Classified: 798
		Underdiagnosed: 150
		Over-diagnosed: 52