Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

One DXA scan limits the apparent accuracy of osteoporosis screening tools but explains at most 13% of what FDA-authorized radiograph devices miss

Yunhao Jiang

biojiang777@gmail.com

University of Edinburgh

Research Article

Keywords: Bone mineral density, Diagnostic accuracy, Dual-energy X-ray absorptiometry, Imperfect reference standard, Opportunistic screening, Artificial intelligence

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11647445/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

One DXA scan limits the apparent accuracy of osteoporosis screening tools but explains at most 13% of what FDA-authorized radiograph devices miss

Yunhao Jiang

School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

Correspondence: biojiang777@gmail.com · ORCID 0009-0004-5954-7315

Mini abstract

Radiograph tools that flag low bone density for DXA referral are judged against one DXA scan with its own error. Even a perfect tool would appear only 70–95% sensitive, yet this explains at most 13% of what FDA-authorized devices miss. A negative device result should not replace DXA.

Abstract

Purpose Opportunistic screening tools, including three radiograph devices authorized by the US Food and Drug Administration (FDA), flag patients with low bone mineral density (BMD) for dual-energy X-ray absorptiometry (DXA), but are judged against one imperfect DXA scan. We asked how accurate a perfect classifier can look against one scan and whether this explains the devices' low sensitivity.

Methods From nationally representative US DXA data we estimated precision-error-free BMD and derived the single-scan frontier: the highest apparent sensitivity any test with errors independent of the scan's can show at each false-positive rate. We validated it in 515 repeat-scan pairs, split device shortfalls into reference and device parts, compared two corrections in 194 simulated scenarios and audited 228 published studies.

Results At national-survey precision, a perfect classifier showed apparent sensitivity of 0.695–0.945 across 24 strata, consistent with all 12 repeat-scan measures. In nine FDA strata, device sensitivity was 0.22–0.79 where the frontier was 0.90–1.00. Per 1,000 people screened, the devices missed 49–420 with DXA-defined low BMD, of whom DXA precision accounted for at most 27 (≤13% of the shortfall at guideline-minimum precision; 17% at national-survey precision). Device results flipped on repetition 2–5 times as often as DXA's would. Conditional-independence correction overstated sensitivity; latent-trait correction was unbiased given known precision and a correct model. Only 13 of 136 full-text studies mentioned DXA precision.

Conclusion One DXA scan makes even a perfect classifier look imperfect but explains little of the authorized devices' low sensitivity, so a negative device result should not replace DXA.

Keywords Bone mineral density · Diagnostic accuracy · Dual-energy X-ray absorptiometry · Imperfect reference standard · Opportunistic screening · Artificial intelligence

Introduction

Osteoporosis is diagnosed when a T-score measured by dual-energy X-ray absorptiometry (DXA) is −2.5 or lower [1]. Opportunistic screening looks for low bone mineral density (BMD) on images that patients already have, such as computed tomography (CT) scans and radiographs, so that they can be referred for DXA; other tools use ultrasound or blood tests. On CT, an automated bone score and vertebral fracture detection predicted fractures and death in 46,219 patients [2]. Three meta-analyses of CT and deep-learning tools report pooled sensitivities of 0.81–0.86 [3–5]. In 2024 the US Food and Drug Administration (FDA) created a device type for software that flags low BMD on radiological images to prompt a clinical assessment of bone health [6], and by October 2026 three radiograph devices had been authorized under this type, with sensitivities of 0.22–0.79 in the nine strata analyzed here [6–8]. For a clinician, sensitivity is the share of patients with low BMD whom the tool flags for

DXA; in six of the nine strata, the devices flagged fewer than half. Almost all these tools are judged against one DXA scan, so a reported sensitivity mixes the tool's error with the reference's.

One scan is not the truth. The International Society for Clinical Densitometry (ISCD) accepts technologist precision errors of up to 2.5% at the femoral neck [1], enough to make labels near the threshold unreliable [9–11]. An imperfect reference biases apparent accuracy [12–14], and forcing a binary label from an uncertain reference adds bias [15]. Standard corrections [16, 17] assume that index-test and reference errors are independent given true status; dependence biases them [14, 18], and dependence arises whenever both tests track bone density. Some radiofrequency echographic multispectrometry studies excluded DXA scans with errors that could not be corrected [19, 20]. A perfect biomarker can look inaccurate against an imperfect reference [21], but no study has measured how far one DXA scan limits the accuracy that any classifier with independent errors can show, or whether this limit explains the modest sensitivity of the authorized devices.

Here we define the single-scan frontier, the highest apparent sensitivity that any test whose errors are independent of the scan's can show at each false-positive rate, and use it to split a test's shortfall into the reference's share and its own. We asked three questions. How accurate can a perfect classifier, one that is right about every person's true BMD status, look against one DXA scan? Does single-scan reference error explain the modest sensitivity of the FDA-authorized devices, or should clinicians take it at face value? Which correction removes the bias, and do published studies report the precision it needs?

Methods

The five analyses, their data sources, the numbers analysed and where each result is reported are shown in Fig. 1.

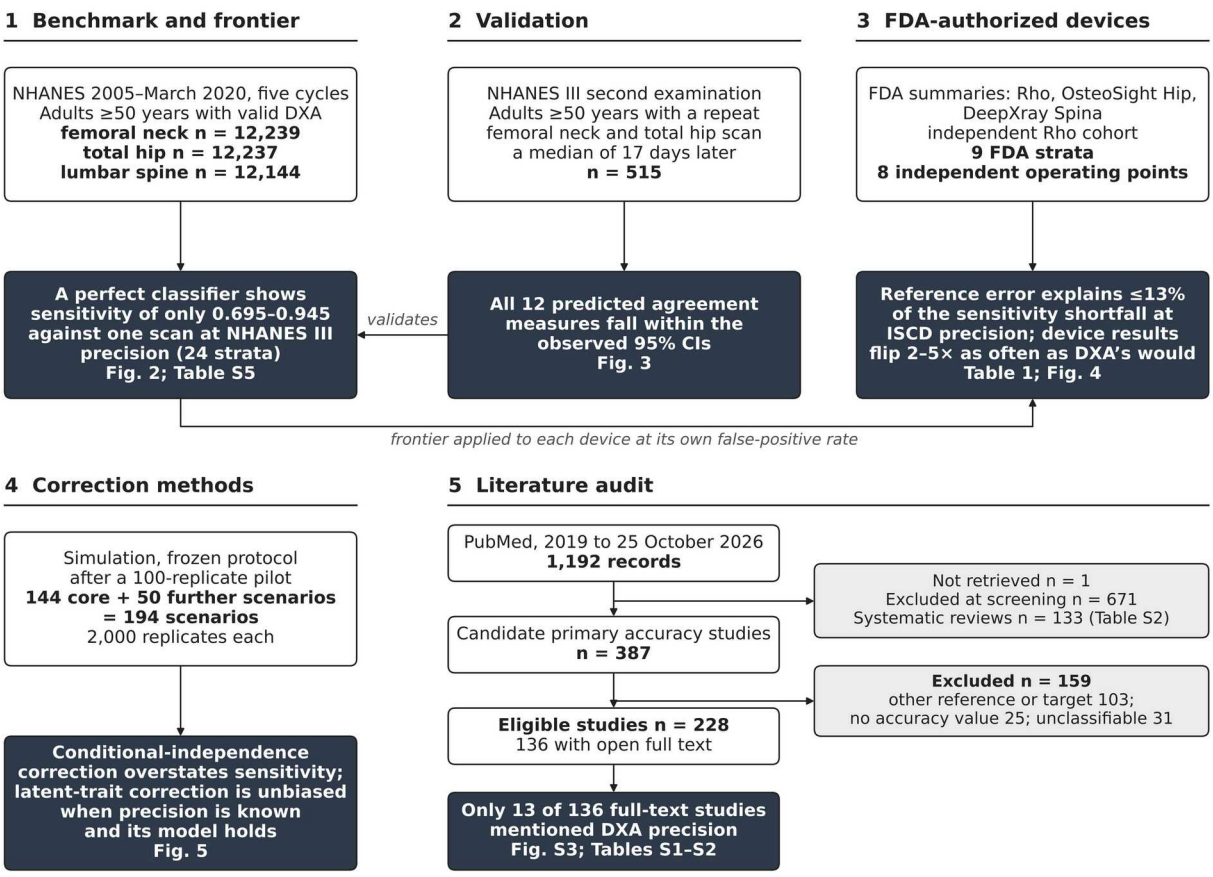


Fig. 1 Study flow. Five analyses, with their data sources and the numbers analysed (white boxes) and the main result of each and where it is reported (dark boxes). FDA strata are device, test cohort and sex; the independent operating points come from a Canadian validation of Rho [22]. Literature exclusions are detailed in Supplementary Methods S2. ISCD, International Society for Clinical Densitometry; NHANES, National Health and Nutrition Examination Survey

Single-scan benchmark and frontier

We modelled one DXA scan as the BMD that the scanner would report without precision error plus a random error proportional to BMD, with standard deviation equal to the coefficient of variation (CV) [23]; repeat pairs confirmed proportional error (Supplementary Methods S1 in Online Resource 1). For the benchmark, osteoporosis was a T-score of −2.5 or lower; as recommended for the femoral neck [24, 25], both sexes were scored against young-adult female reference values (Supplementary Methods S1 [26, 27]). A perfect classifier (X =

D in the figures) reports each person's error-free status D. Against one scan its apparent sensitivity falls below 1, because some people whose true BMD lies just above the threshold scan just below it: the perfect classifier calls them negative, and the scan counts this as a missed case. This apparent sensitivity and specificity are the benchmark. Ranking people by their error-free BMD traces the single-scan frontier: at each false-positive rate, no test whose errors are independent of the scan's can show a higher apparent sensitivity (Neyman–Pearson lemma; formal model in Supplementary Methods S1). For a single-site reference the perfect classifier lies on this frontier, so its benchmark is the highest sensitivity attainable at its own false-positive rate. The area under the frontier, the AUC ceiling, is the largest apparent area under the receiver operating characteristic curve (AUC) that such a test can show. Frontier sensitivity at false-positive rates of 1%, 5% and 10% is given in Table S5.

True BMD distribution and precision scenarios

We fitted survey-weighted Gaussian mixtures to log BMD in adults aged ≥50 years from five National Health and Nutrition Examination Survey (NHANES) cycles (2005–March 2020; femoral neck n = 12,239, total hip 12,237, lumbar spine 12,144) by site, sex and age group. Extreme deconvolution with known error variance [28] gave the distribution without precision error, and fits with 1–4 components were averaged by Bayesian information criterion. Two hundred Rao–Wu bootstrap replicates gave 95% confidence intervals (CI) for the headline femoral-neck benchmarks [29]. Other populations were represented by shifting the distribution to reproduce their DXA-defined prevalence (Supplementary Methods S1). Precision came from NHANES III repeat scans (1988–1994; femoral neck 3.01%, total hip 2.40%), close to the 3.2% reported by that survey's quality-control program [30], with the spine assumed at 1.9%. The ISCD minimum acceptable precision (2.5%/1.8%/1.9%) [1] was the primary scenario for the devices; the other scenarios are listed in Table S3.

Validation against observed repeat scans

In the NHANES III second-examination subsample, 515 adults aged ≥50 years had a repeat hip scan a median of 17 days after the first. With the first scan as the reference and the repeat as the index test, we compared observed agreement with model predictions made with cross-fitted precision. The prestated criterion was that predictions fall inside the observed 95% CIs; the full 2×2 table was tested for goodness of fit. A companion study on diagnostic confirmation uses the same repeat examinations.

FDA-authorized devices

We took accuracy and prevalence from the FDA summaries of the three authorized devices, Rho (16 Bit Inc., Toronto, Canada; De Novo, DEN230023), OsteoSight Hip (Naitive Technologies Ltd, Cambridge, UK; 510(k), K251408) and DeepXray Spina (Alpha Intelligence Manifolds, Inc., New Taipei City, Taiwan; 510(k), K253192) [6–8] and from an independent validation of Rho [22], and repeatability from the Rho and DeepXray Spina summaries [6, 8]. These sources gave nine FDA strata (device, FDA test cohort and sex) and eight independent operating points, that is, reported sensitivity–specificity pairs (Tables S6–S7).

For each, the sex-specific BMD distribution was shifted to the reported DXA-defined prevalence and the reference rule was reproduced as reported: femoral neck $T < -1.0$, L1–L4 $T \leq -2.5$, or the lowest of femoral neck and L1–L4 $T < -1.0$. For lowest-of-sites rules, the frontier orders people by how likely one scan is to call them positive, given their true BMD at each site (Supplementary Methods S4). The reference share is the part of a device's sensitivity shortfall that one scan's precision error accounts for: (1 − frontier sensitivity)/(1 − device sensitivity), with the frontier read at the device's own false-positive rate. Ranges span the corners of the reported CIs. Latent-trait corrections used effective sample sizes derived from the reported CIs. Device label changes on repeat imaging were compared with the model

rate at which two independent DXA scans disagree at the same threshold and DXA-positive proportion.

Corrections and simulation

We compared naive (uncorrected) accuracy against one scan, the conditional-independence correction [16] supplied with the model's reference accuracy and prevalence, and a latent-trait correction related to random-effects latent-class models [31]. The latent-trait correction treats the index test as a noisy reading of the same true BMD: from apparent sensitivity and specificity, the BMD distribution and DXA precision, it estimates the reading's noise and threshold and returns true accuracy with delta-method or bootstrap intervals. After a 100-replicate pilot, the frozen protocol crossed population, CV, true AUC, sample size and design in 144 core scenarios; sensitivity analyses and an index-misspecification amendment brought the total to 194, each with 2,000 replicates. A correction was adequate if $|\text{bias}| \leq 0.01$ and 95% interval coverage was 0.925–0.975 at $n \geq 500$ (Supplementary Methods S1). A post hoc extension compared six bone reference rules, from one scan to selective re-scanning and confirmation of positive results, under a fixed scan budget (Supplementary Methods S3).

Literature audit

We searched PubMed for accuracy studies of osteoporosis classifiers against DXA-defined osteoporosis published from 2019 to 25 October 2026 (Supplementary Methods S2). Reported operating points were compared with prevalence-matched benchmarks and frontiers, and reported sensitivity was corrected at ISCD precision (Table S1). Analyses used Python 3 (NumPy, SciPy, pandas).

Results

Against one scan, even a perfect classifier looks imperfect, and repeat-scan agreement was consistent with the model

Against one scan at NHANES III precision, a perfect classifier would show femoral-neck sensitivity of only 0.872 in women and 0.777 in men aged ≥50 years. One scan mislabeled 0.9% of women as negative and 1.2% as positive, almost all within 0.5 T-score units of the threshold (99.98%; Fig. 2a). Across 24 site–sex–age strata, benchmark sensitivity was 0.695–0.945, or 0.777–0.945 at ISCD precision, at specificities of at least 0.971 (Fig. 2d–f; Table S12). Higher sensitivity is available only at a higher false-positive rate: at 1%, frontier sensitivity was 0.833–1.000 (Table S5), and the AUC ceiling was at least 0.992. The benchmark rose with prevalence, from 0.863 at the US value of 5.5% to 0.924 at 33% and 0.970 at 72% (Fig. 2c). For every 1,000 women aged ≥65 years screened at the femoral neck, one scan at ISCD precision labels 14 as osteoporotic who are not and misses 12 who are.

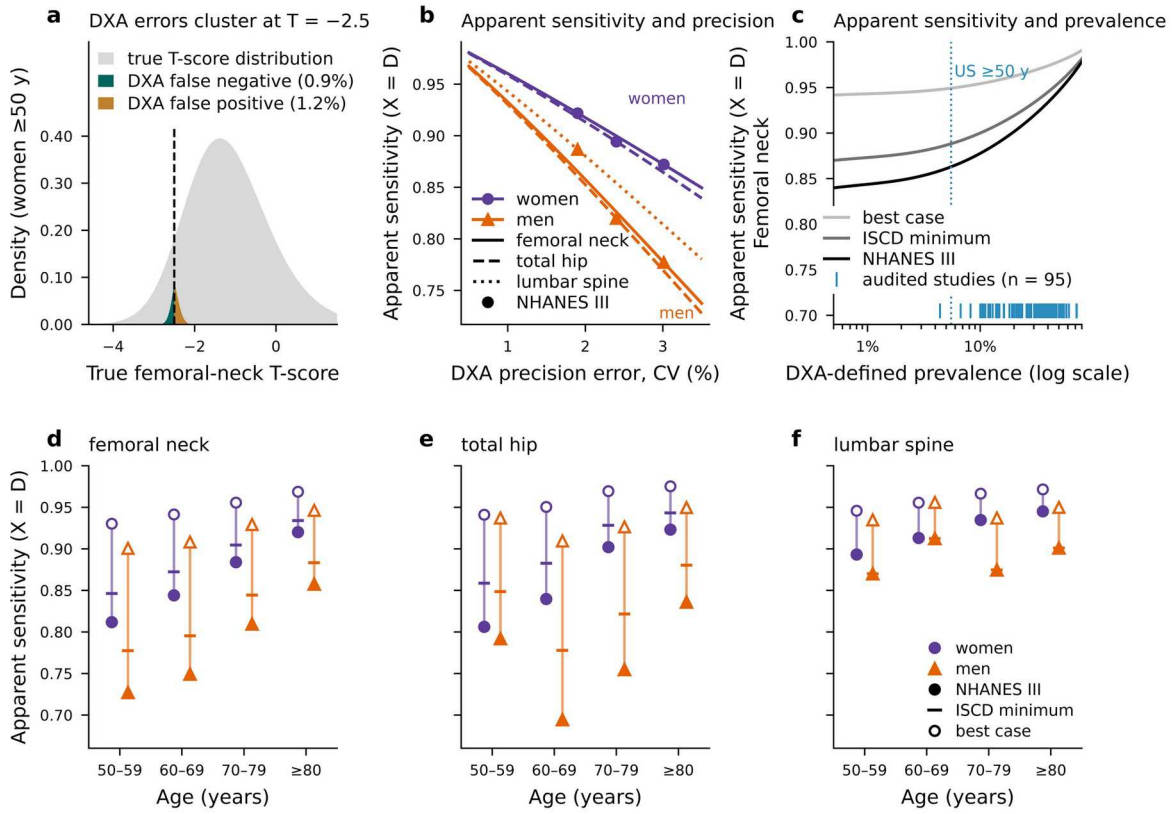


Fig. 2 Precision error in one DXA scan mislabels people only near the threshold, yet it lowers the apparent sensitivity of a perfect classifier, most at poor precision and low prevalence. (a) Distribution of precision-error-free femoral-neck T-score in US women aged ≥50 years (five NHANES cycles, 2005–March 2020), with the shares mislabeled by one scan at NHANES III precision (CV 3.01%). (b) Apparent sensitivity of a perfect classifier (X = D) against one scan by precision error, site and sex (adults ≥50 years); symbols mark NHANES III precision (femoral neck 3.01%, total hip 2.40%, assumed spine 1.9%). In women the spine and femoral-neck curves overlap. (c) The femoral-neck benchmark (both sexes) by DXA-defined prevalence, obtained by shifting the BMD distribution; ticks mark 95 audited studies outside case–control designs (two, at 84% and 87%, lie beyond the axis); dotted line, US DXA-defined prevalence. (d–f) Apparent sensitivity across 24 site–sex–age strata at the femoral neck (d), total hip (e) and lumbar spine (f) in US adults aged ≥50 years: filled symbols, NHANES III precision (range 0.695–0.945); bars, ISCD minimum acceptable precision (0.777–0.945); open symbols, assumed best case. For the spine both scenarios assume 1.9%.

Women, purple circles; men, orange triangles. Frontier sensitivity at fixed false-positive rates is given in Table S5. Each benchmark is the highest sensitivity attainable at its own false-positive rate. ISCD, International Society for Clinical Densitometry

In validation against observed repeat scans, all 12 predicted agreement measures between repeat and first scans fell inside the observed 95% CIs at both hip sites (Fig. 3). At the femoral neck, the full 2×2 table rejected the whole-survey BMD distribution (goodness of fit $p = 0.009$) but not the distribution fitted to the subsample ($p = 0.80$), whose DXA-defined prevalence was lower (5.6% versus 11.5%). Full-table predictions therefore need a prevalence-matched distribution.

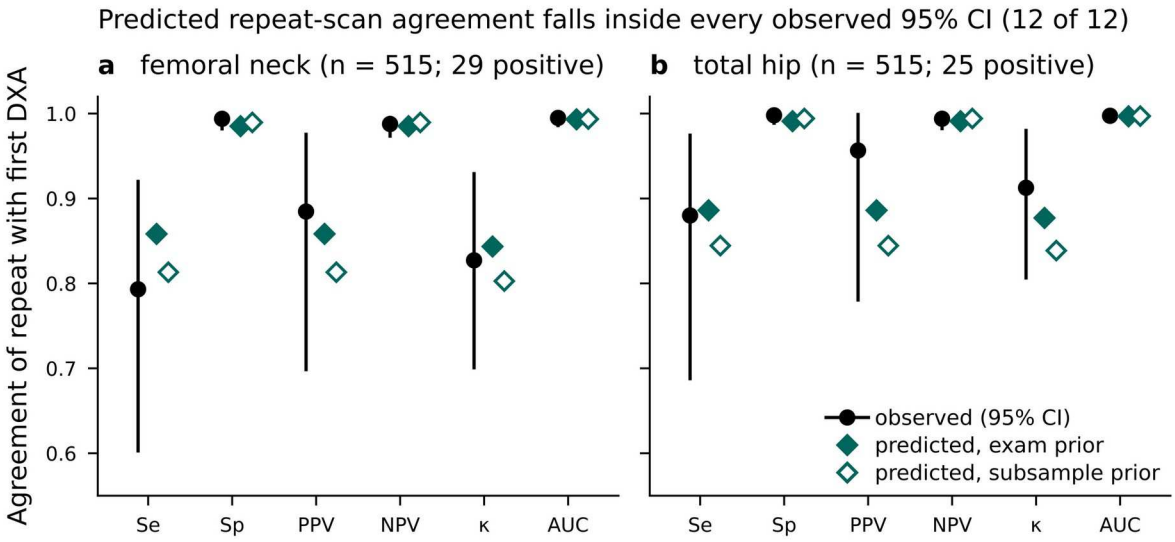


Fig. 3 Observed repeat-scan agreement was consistent with the model behind the frontier: predictions fell inside all 12 observed 95% confidence intervals. Agreement between repeat and first DXA scans in 515 adults aged ≥50 years from the NHANES III second-examination subsample. Error bars, 95% CIs (Clopper–Pearson; bootstrap for κ; DeLong for AUC). Exam prior, BMD distribution fitted to the full NHANES III examination and standardized to the subsample's sex–age mix; subsample prior, distribution fitted to first scans. Predictions use cross-fitted precision and concern a second noisy scan, not a perfect classifier; sensitivity and specificity use the first-scan label as reference. NPV, negative predictive value; Se, sensitivity; Sp, specificity

Against the frontier, reference error explains at most 13% of the authorized devices' shortfall

In the nine FDA strata, device sensitivity was 0.22–0.79, whereas the frontier at the device's own false-positive rate was 0.90–1.00 (Table 1; Fig. 4a). Per 1,000 patients screened, the devices left 49–420 patients with DXA-defined low BMD or osteoporosis unflagged, of whom the scan's imprecision accounts for 0–27 (Table 1). In the OsteoSight Hip cohort, the device did not flag 420 of the 667 with DXA-defined low BMD, whereas a test reading error-free BMD, set to the device's false-positive rate, would have appeared to miss 27. Reference error therefore accounted for 0.3–12.5% of the sensitivity shortfall (median 4.1%; 0.3–16.9% at NHANES III precision), and the device's own error cost at least 0.20 of sensitivity in every stratum (Fig. 4c; Fig. S1). In the independent cohort the share was 0–13.3% (median 2.8%; Table S7). It was largest at strict operating points in men: 12.5% for Rho in OAI-cohort men at specificity 0.98 (6.0–21.1% across the reported CIs), and 13.3% in independent-cohort men at Rho score cutoff 8 (specificity 0.992).

Table 1 FDA-authorized screening devices against the single-scan frontier (ISCD minimum precision)

Device (route; submission)	Cohort, stratum	Reference definition	Prevalence (%)	Reported Se / Sp	Benchmark Se / Sp	Frontier Se at device FPR	Reference share, % (range)	Missed per 1,000: device / attributable to DXA precision
OsteoSight Hip (510(k); K251408)	6 US cohorts, all	FN T < −1.0	66.7	0.370 / 0.951	0.970 / 0.935	0.960	6.3 (4.7–8.2)	420 / 27
DeepXray Spina (510(k); K253192)	3 test sets, women	L1–L4 T ≤ −2.5	27.5	0.787 / 0.934	0.943 / 0.980	0.991	4.1 (1.7–10.9)	59 / 2
	3 test sets, men	L1–L4 T ≤ −2.5	11.4	0.571 / 0.942	0.921 / 0.991	0.999	0.3 (0.0–6.8)	49 / 0
Rho (De Novo; DEN230023)	OAI, women	FN T < −1.0	51	0.44 / 0.90	0.955 / 0.953	0.984	2.8 (0.9–6.8)	286 / 8
	OAI, men	FN T < −1.0	28	0.22 / 0.98	0.923 / 0.973	0.902	12.5 (6.0–21.1)	218 / 27
	TNI6, women	FN or L1–L4 T < −1.0	89	0.70 / 0.90	0.989 / 0.902	0.989	3.6 (2.7–4.8)	267 / 10

Device (route; submission)	Cohort, stratum	Reference definition	Prevalence (%)	Reported Se / Sp	Benchmark Se / Sp	Frontier Se at device FPR	Reference share, % (range)	Missed per 1,000: device / attributable to DXA precision
	TNI6, men	FN or L1–L4 T < −1.0	66	0.49 / 0.90	0.966 / 0.938	0.981	3.7 (1.9–6.4)	337 / 13
	OMN, women	FN or L1–L4 T < −1.0	69	0.48 / 0.94	0.971 / 0.938	0.970	5.8 (2.3–14.3)	359 / 21
	OMN, men	FN or L1–L4 T < −1.0	52	0.41 / 0.92	0.953 / 0.954	0.976	4.1 (1.7–10.4)	307 / 12

Se, sensitivity; Sp, specificity; FPR, false-positive rate; FN, femoral neck. Benchmark, apparent accuracy of a perfect classifier against one scan. Frontier, the highest apparent sensitivity that any test with errors independent of the scan's can show at the device's false-positive rate. Reference share, (1 − frontier Se)/(1 − device Se): the part of the device's sensitivity shortfall attributable to single-scan precision error; ranges span the corners of the reported confidence intervals. For TNI6 and OMN (lowest of FN or L1–L4) the frontier is the posterior frontier of the two-site rule (Supplementary Methods S4). OsteoSight Hip: full intended-use population, no result counted negative; with a device result (n 2,596) the share is 6.2%. DeepXray Spina: image-level accuracy; pooled sexes, 3.2%. Rho prevalences are reported to whole percentages. NHANES III and best-case shares are in Table S6, the independent cohort in Table S7. Missed per 1,000 screened: 1,000 × prevalence × (1 − device Se); attributable to DXA precision: 1,000 × prevalence × (1 − frontier Se), from the tabulated values; if device and DXA errors were correlated, this would be an upper bound.

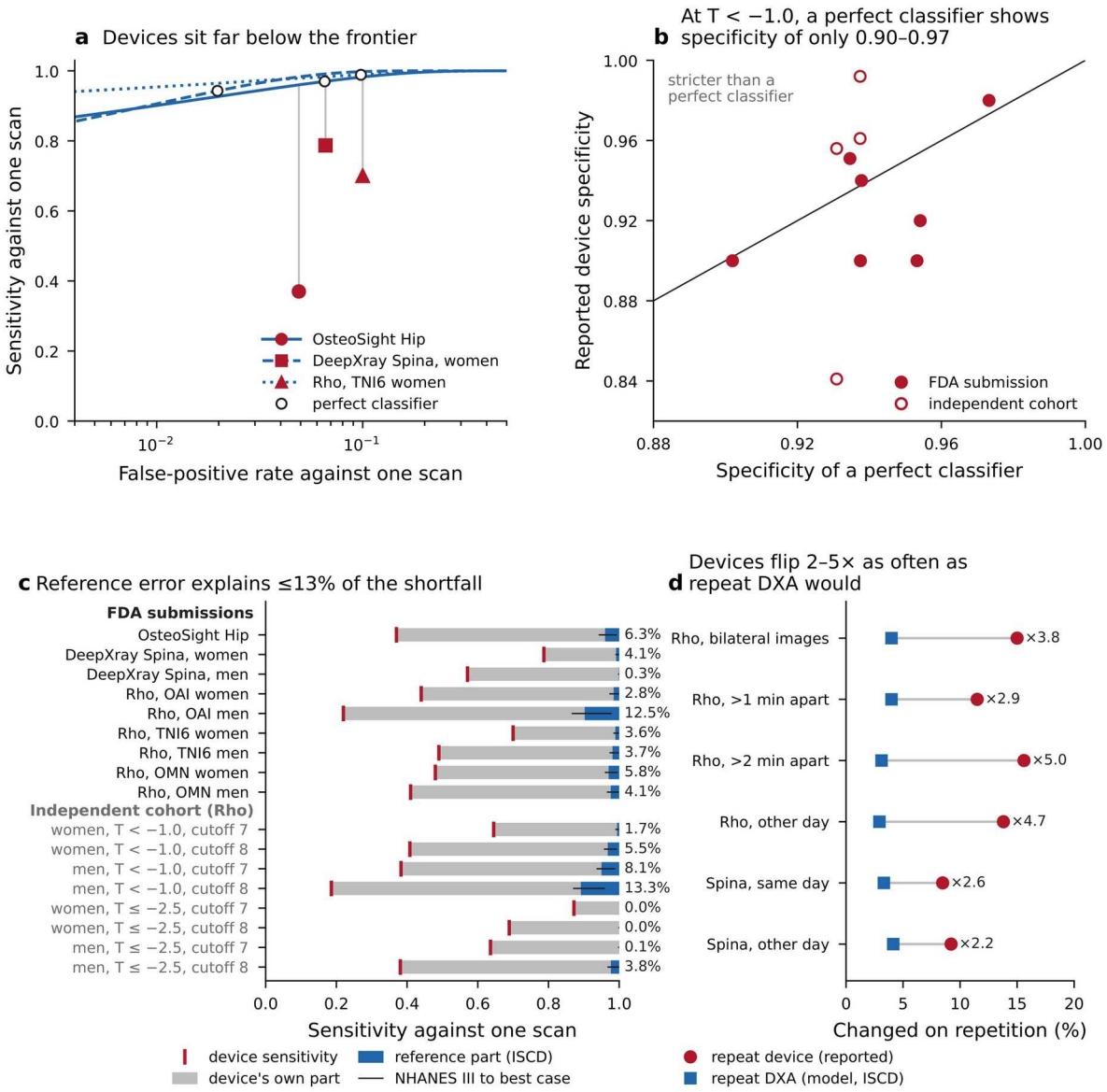


Fig. 4 FDA-authorized devices sit far below the frontier that their single-scan reference allows. (a) Each authorized device lies far below the frontier its own reference allows: frontier at ISCD precision for OsteoSight Hip, DeepXray Spina (women) and Rho (TNI6 women; posterior frontier of the two-site rule), with the device's reported operating point (red) and the perfect classifier's point (open circle); gray lines mark the gap at the device's false-positive rate. (b) At T < −1.0, a perfect classifier would show specificity of only 0.90–0.97 against one scan, and 6 of 11 device operating points reported more (seven FDA strata, filled; four independent-cohort points, open), with each cohort's BMD distribution shifted to its prevalence. Points above the identity line use thresholds stricter than a perfect classifier. (c)

Single-scan reference error explains 0–13% of the sensitivity shortfall: device sensitivity (red tick), the device's own part (gray) and the reference part at ISCD precision (blue) for each operating point, with the range from NHANES III to best-case precision (black line); labels give the reference share. The independent cohort is Rho in a Canadian validation study at score cutoff 7 or 8 [22]. (d) Repeating the device changed its result 2–5 times as often as repeating the DXA scan would: reported device flip rates (circles) against the model rate for two independent DXA scans (squares) at the same threshold and DXA-positive proportion (ISCD precision). OAI, OMN and TNI6 are the cohorts named in the FDA De Novo summary Two further results point to the devices rather than to the scan's precision. Repeating the device changed its result in 8.5–15.6% of patients, 2.2–5.0 times the rate at which two DXA scans would disagree in the same populations (2.9–4.1%), and 2.6–5.0 times in same-day repeats (Fig. 4d; Table S9). And against one scan at $T < -1.0$, with each cohort's BMD distribution shifted to its prevalence, a perfect classifier would show specificity of only 0.90–0.97, yet 6 of 11 device operating points reported more (0.94–0.99; Fig. 4b): these are strict thresholds, not near-perfect agreement with true status. Latent-trait correction raised reported sensitivity by less than 0.01 and specificity by at most 0.003 (Table S8). AUC ceilings were 0.993–0.998, against device AUCs of 0.79–0.957, and 0.959 for DeepXray Spina in an independent cohort [32].

Conditional-independence correction overstated sensitivity; latent-trait correction worked when the precision was known

The conditional-independence correction overestimated true sensitivity in all 144 core scenarios (bias +0.014 to +0.112; above 0.02 in 134) and in all 194 scenarios overall, as the proof in Supplementary Methods S1 requires, whereas the naive estimate was off by more than 0.02 in only 17 core scenarios (Fig. 5a, b). Latent-trait correction met the adequacy criteria in all 108 core scenarios with $n \geq 500$ (bias −0.003 to +0.007; coverage 0.940–0.974)

and stayed within 0.009 of the truth across the nuisance, site and operating-point variations. Assuming a CV that was wrong by one percentage point biased the corrected sensitivity by 0.0067–0.0085, about as much as the naive bias the correction removes (0.0076–0.0153 in the same scenarios). When half the index response depended only on binary status, latent-trait bias was −0.032 to −0.020 (Fig. 5c).

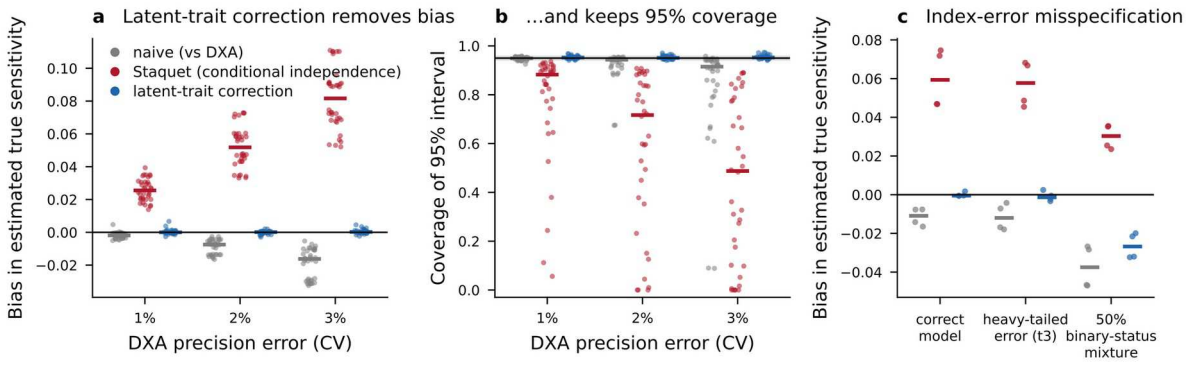


Fig. 5 Conditional-independence correction overstated sensitivity in every scenario, whereas latent-trait correction removed the bias when its model was correct. Simulations with 2,000 replicates per scenario; in each group, naive, conditional-independence and latent-trait estimates are plotted left to right. (a, b) Bias and coverage in the 108 core scenarios with n ≥500 by DXA precision; each dot is one scenario, bars are medians; gray band, Monte Carlo error around the nominal 0.95. The conditional-independence (Staquet) correction was given the true reference accuracy. (c) Bias with correctly specified and misspecified index mechanisms (femoral neck, women 60–69 years, true AUC 0.85, consecutive sampling, CV 2–3%, n 500–1,000; four scenarios per group). The mixed mechanism generates half the index results from a continuous-trait response and half from fixed error rates within binary latent-status classes

How the reference is built also changes what a study reports. Confirming only positive scans turned the one-scan underestimate of sensitivity (−0.7 to −3.0 percentage points) into an overestimate (+1.2 to +2.2) in all four bone scenarios, and after latent-trait correction one scan

per participant was the most efficient reference tested (Supplementary Methods S3; Table S4; Fig. S2).

Few published studies report the precision needed to apply the frontier

The search returned 1,192 PubMed records. Of 387 candidate primary studies, 228 met the criteria. Only 13 studies (6%) mentioned DXA precision, all among the 136 with open full text (9.6%), and 10 reported a value; one repeated scans to measure precision, and none used repeat scans to form the reference. Of 29 reported operating points, 28 lay at or below their matched frontier. The other (12 of 13 at specificity 0.986) lay above it but below the most generous frontier (Fig. S3). Latent-trait correction raised published sensitivity by a median of 0.004 and the conditional-independence correction by 0.039; for six meta-analyses the corrections were 0.0004–0.014 and 0.007–0.084 (Table S2).

Discussion

In this analysis of nationally representative DXA data, regulatory summaries and published studies, even a perfect classifier looked imperfect against one DXA scan, and no test whose errors are independent of the scan's can rise above the single-scan frontier. For the FDA-authorized devices, the reference's share of the shortfall was at most 13% at ISCD precision, correcting for it changed their reported accuracy by less than 0.01, and their results flipped on repetition 2–5 times as often as a repeat DXA scan's would. A repeated or more precise scan would not rescue their low sensitivity. Conditional-independence correction overstated sensitivity in every simulated scenario, latent-trait correction removed the bias when the scanner's precision was known and its model held, and only 13 of 136 published studies with full text mentioned that precision.

Earlier work quantified how precision error and cross-calibration misclassify individual scans [9–11], showed that conditional dependence biases corrections [18] and that latent-class

corrections fail when the dependence mechanism is misspecified [33], bounded true accuracy when the reference's error rates are known [34], and showed that how a reference label is formed changes estimated accuracy [15, 35]. We instead derive a bound on apparent accuracy from the scanner's measured precision, for any classifier with independent errors, test the corrections against it and apply it to devices already on the market.

Clinical implications

For clinicians, the result has a direct reading. The indications for use of all three devices state that they should not be used to rule out low BMD [6–8], and our results show that this caution reflects the devices' own error, not the imprecision of their DXA reference: a negative or absent device report should not delay DXA for a patient who otherwise qualifies. In practice, OsteoSight Hip, with a sensitivity of 0.37 in its full intended-use population, left about six in ten patients with low BMD unflagged, and the scan's imprecision explains little of that. For individual patients, almost all labels that one scan gets wrong lie within 0.5 T-score units of −2.5 (Fig. 2a). Conversely, against one femoral-neck scan in adults aged ≥50 years, a reported sensitivity of 0.86–0.89 with specificity above 0.99 is what a perfect classifier would show at NHANES III or ISCD precision (Table S12).

For regulators and developers, the frontier gives single-scan performance a yardstick. The special controls for this device type already require a clinically justified ground truth and testing of the device's reproducibility [6], but not the reproducibility of the reference. Yet a repeat DXA scan would change 2.4–6.5% of reference labels in the evaluation populations. Device evaluations should therefore report the facility's DXA precision; the threshold, site rule and prevalence by sex; accuracy against the matched benchmark and frontier; how often repeating the reference and repeating the device change the result, side by side; a reference rule without positive-only confirmation; and the measured CV used for any correction (Table

S13). STARD asks for the reference standard and its cutoff [36]; these items add what a DXA reference needs.

For accuracy studies, the reference rule and the estimator should be chosen together. Both depend on the scanner's precision. With latent-trait correction, scan everyone once and spend repeat scans on a precision study in the study's own scanner: a CV that is one point wrong biases corrected sensitivity by about 0.8 points, whereas a 30-pair precision study leaves about 0.3 points of uncertainty (Supplementary Methods S1). Without correction, avoid both universal duplication and confirmation of positive scans only (Supplementary Methods S3), which turned underestimates into overestimates and, like excluding uncertain results [15] or discrepant analysis [37], can inflate apparent sensitivity. Conditional-independence correction should not be used with a DXA reference when the index test tracks bone density, because both tests then err in the same people near the threshold.

Strengths and limitations

This study has several strengths. The BMD distributions come from five nationally representative survey cycles, the model was checked against observed repeat scans, and the simulation protocol was frozen after a pilot and before the main runs. It also has limitations. The model target is DXA BMD without precision error, not bone mass. Soft-tissue accuracy errors [38], which persist on repeat scans, count here as device error; analysis errors, cross-calibration bias [11] and BMD change before the scan would add to the reference share. The device frontier assumes that device and DXA errors are independent given true BMD; errors shared through soft tissue or positioning would let a device agree with the scan's own errors, so the reference shares would then be upper bounds. Shifting the US distributions to each cohort's prevalence keeps their shape; matching OsteoSight's reported spread as well lowered its share from 6.3% to 4.7% (Table S10). One CV per site was applied to multi-facility cohorts, and the ISCD and NHANES III scenarios bracket that CV. Device inputs were

published summaries, and two sources reported image-level accuracy. The repeat-scan check had few positive participants, and the simulations assumed known nuisance parameters.

Conclusion

Against one DXA scan, even a perfect osteoporosis classifier looks imperfect. For the FDA-authorized radiograph devices this limit explains at most 13% of what they miss at ISCD precision: their low sensitivity is not a scan-precision artifact. Only 13 of 136 published studies mentioned the scanner's precision; accuracy studies and device evaluations should report it, read results against the matched frontier, and use latent-trait rather than conditional-independence correction. The devices may prompt DXA for some patients, but a negative result should not replace DXA when it is indicated.

Declarations

Funding No funding was received for conducting this study.

Conflicts of interest The author has no relevant financial or non-financial interests to disclose.

Ethics approval This study used only publicly available, de-identified data from NHANES, collected under protocols approved by the National Center for Health Statistics ethics review bodies, and published regulatory and study summaries. This article does not contain any studies with human participants or animals performed by the author.

Consent to participate For this type of study formal consent is not required.

Consent for publication Not applicable.

Data and code availability NHANES data are publicly available from the National Center for Health Statistics, and the FDA summaries are public [6–8]. Analysis code, fitted BMD distributions and result tables, including the precision-misspecification simulations (sim_sensitivity_S3.csv), the frontier table (Table S5) and the device analysis (Tables S6–

S10), are available at https://zenodo.org/records/23851713?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjRjMTIwNjA0LTQ3MWUtNGQ4Yy1iZjc5LTAyZDcxZWIyYTcyMSIsImRhdGEiOnt9LCJyYW5kb20iOiI4Mzg2ZjVkNjQyNDUyNzYyZmVhMzdkYzk3ZDljMTRmZSJ9.BeJWSm-hkj7zXTjDQiPKlwox8h8-jRB3a0sQPXwPckjVfOuCXT_-9tDHxy3BmBxKc7Uh9slx4vRvPBGCJY3X1w and will be published openly on Zenodo (https://doi.org/10.5281/zenodo.23851713) on journal publication.

Supplementary information Online Resource 1, appended to this preprint, contains Supplementary Methods S1–S4, Tables S1–S13 and Figs. S1–S3.

References

1. International Society for Clinical Densitometry (2023) 2023 ISCD Official Adult Positions. https://iscd.org/text-based-official-adult-positions/. Accessed 31 October 2026

2. Roux C, Rozes A, Ajrouche A et al (2026) Fractures and mortality prediction by opportunistic screening of osteoporosis. Osteoporos Int. https://doi.org/10.1007/s00198-026-08270-5

3. Zhu Y, Triphuridet N, Yip R et al (2021) Opportunistic CT screening of osteoporosis on thoracic and lumbar spine: a meta-analysis. Clin Imaging 80:382–390. https://doi.org/10.1016/j.clinimag.2021.08.005

4. Yen TY, Ho CS, Chen YP, Pei YC (2024) Diagnostic accuracy of deep learning for the prediction of osteoporosis using plain X-rays: a systematic review and meta-analysis. Diagnostics (Basel) 14(2):207. https://doi.org/10.3390/diagnostics14020207

5. Amani F, Amanzadeh M, Hamedan M, Amani P (2024) Diagnostic accuracy of deep learning in prediction of osteoporosis: a systematic review and meta-analysis. BMC Musculoskelet Disord 25(1):991. https://doi.org/10.1186/s12891-024-08120-7

6. US Food and Drug Administration (2024) De Novo classification request DEN230023 for Rho (16 Bit Inc.), establishing 21 CFR 892.1171. De Novo summary, granted 9 April 2024. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN230023.pdf. Accessed 31 October 2026

7. US Food and Drug Administration (2025) 510(k) premarket notification K251408: OsteoSight Hip (Naitive Technologies Ltd). Summary, cleared 2 September 2025. https://www.accessdata.fda.gov/cdrh_docs/pdf25/K251408.pdf. Accessed 31 October 2026

8. US Food and Drug Administration (2026) 510(k) premarket notification K253192: DeepXray Spina (Alpha Intelligence Manifolds, Inc.). Summary, cleared 10 June 2026. https://www.accessdata.fda.gov/cdrh_docs/pdf25/K253192.pdf. Accessed 31 October 2026

9. Phillipov G, Seaborn CJ, Phillips PJ (2001) Reproducibility of DXA: potential impact on serial measurements and misclassification of osteoporosis. Osteoporos Int 12(1):49–54. https://doi.org/10.1007/s001980170157

10. Kiebzak GM, Faulkner KG, Wacker W, Hamdy R, Seier E, Watts NB (2007) Effect of precision error on T-scores and the diagnostic classification of bone status. J Clin Densitom 10(3):239–243. https://doi.org/10.1016/j.jocd.2007.03.002

11. Fahrenholtz SJ, Long JR, Whitaker MD, Sensakovic WF (2023) Quantifying misdiagnosis rates from cross-calibration biases and precision errors in dual-energy X-ray absorptiometry of the femoral neck. Med Phys 50(3):1623–1634. https://doi.org/10.1002/mp.16057

12. Boyko EJ, Alderman BW, Baron AE (1988) Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. J Gen Intern Med 3(5):476–481. https://doi.org/10.1007/BF02595925

13. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM (2009) A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 62(8):797–806. https://doi.org/10.1016/j.jclinepi.2009.02.005

14. Umemneku Chikere CM, Wilson KJ, Allen AJ, Vale L (2021) Comparative diagnostic accuracy studies with an imperfect reference standard – a comparison of correction methods. BMC Med Res Methodol 21(1):67. https://doi.org/10.1186/s12874-021-01255-4

15. Jenniskens K, Naaktgeboren CA, Reitsma JB, Hooft L, Moons KGM, van Smeden M (2019) Forcing dichotomous disease classification from reference standards leads to bias in diagnostic accuracy estimates: a simulation study. J Clin Epidemiol 111:1–10. https://doi.org/10.1016/j.jclinepi.2019.03.002

16. Staquet M, Rozencweig M, Lee YJ, Muggia FM (1981) Methodology for the assessment of new dichotomous diagnostic tests. J Chronic Dis 34(12):599–610. https://doi.org/10.1016/0021-9681(81)90059-x

17. Hui SL, Walter SD (1980) Estimating the error rates of diagnostic tests. Biometrics 36(1):167–171. https://doi.org/10.2307/2530508

18. Vacek PM (1985) The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 41(4):959–968. https://doi.org/10.2307/2530967

19. Di Paola M, Gatti D, Viapiana O et al (2019) Radiofrequency echographic multispectrometry compared with dual X-ray absorptiometry for osteoporosis diagnosis on lumbar spine and femoral neck. Osteoporos Int 30(2):391–402. https://doi.org/10.1007/s00198-018-4686-3

20. Cortet B, Dennison E, Diez-Perez A et al (2021) Radiofrequency echographic multi spectrometry (REMS) for the diagnosis of osteoporosis in a European multicenter clinical context. Bone 143:115786. https://doi.org/10.1016/j.bone.2020.115786

21. Waikar SS, Betensky RA, Emerson SC, Bonventre JV (2012) Imperfect gold standards for kidney injury biomarker evaluation. J Am Soc Nephrol 23(1):13–21. https://doi.org/10.1681/ASN.2010111124

22. Auriat AM, Patel M, Zhang L et al (2026) Evaluating an artificial intelligence software for opportunistic low bone mineral density and osteoporosis screening: a validation study. JBMR Plus 10(2):ziaf191. https://doi.org/10.1093/jbmrpl/ziaf191

23. Glüer CC, Blake G, Lu Y, Blunt BA, Jergas M, Genant HK (1995) Accurate assessment of precision errors: how to measure the reproducibility of bone densitometry techniques. Osteoporos Int 5(4):262–270. https://doi.org/10.1007/BF01774016

24. Kanis JA, McCloskey EV, Johansson H, Oden A, Melton LJ 3rd, Khaltaev N (2008) A reference standard for the description of osteoporosis. Bone 42(3):467–475. https://doi.org/10.1016/j.bone.2007.11.001

25. Kanis JA, Adachi JD, Cooper C et al (2013) Standardising the descriptive epidemiology of osteoporosis: recommendations from the Epidemiology and Quality of Life Working Group of IOF. Osteoporos Int 24(11):2763–2764. https://doi.org/10.1007/s00198-013-2413-7

26. Looker AC, Wahner HW, Dunn WL et al (1998) Updated data on proximal femur bone mineral levels of US adults. Osteoporos Int 8(5):468–489. https://doi.org/10.1007/s001980050093

27. Melamed A, Vittinghoff E, Sriram U, Schwartz AV, Kanaya AM (2010) BMD reference standards among South Asians in the United States. J Clin Densitom 13(4):379–384. https://doi.org/10.1016/j.jocd.2010.05.007

28. Bovy J, Hogg DW, Roweis ST (2011) Extreme deconvolution: inferring complete distribution functions from noisy, heterogeneous and incomplete observations. Ann Appl Stat 5(2B):1657–1677. https://doi.org/10.1214/10-AOAS439

29. Rao JNK, Wu CFJ (1988) Resampling inference with complex survey data. J Am Stat Assoc 83(401):231–241. https://doi.org/10.1080/01621459.1988.10478591

30. Wahner HW, Looker A, Dunn WL, Walters LC, Hauser MF, Novak C (1994) Quality control of bone densitometry in a national health survey (NHANES III) using three mobile

examination centers. J Bone Miner Res 9(6):951–960. https://doi.org/10.1002/jbmr.5650090621

31. Qu Y, Tan M, Kutner MH (1996) Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 52(3):797–810. https://doi.org/10.2307/2533043

32. Tseng TH, Huang TT, Huang JE et al (2026) AI-derived bone mineral density from standard radiographs compared with DXA for fracture prediction in a 10-year real-world cohort study. Osteoporos Int 37(9):2355–2364. https://doi.org/10.1007/s00198-026-08100-8

33. Albert PS, Dodd LE (2004) A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics 60(2):427–435. https://doi.org/10.1111/j.0006-341X.2004.00187.x

34. Emerson SC, Waikar SS, Fuentes C, Bonventre JV, Betensky RA (2018) Biomarker validation with an imperfect reference: issues and bounds. Stat Methods Med Res 27(10):2933–2945. https://doi.org/10.1177/0962280216689806

35. Vach W (2026) Conditions for unbiased estimation of sensitivity and specificity when using a probabilistic reference standard could be identified. J Clin Epidemiol 196:112308. https://doi.org/10.1016/j.jclinepi.2026.112308

36. Bossuyt PM, Reitsma JB, Bruns DE et al (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 351:h5527. https://doi.org/10.1136/bmj.h5527

37. Miller WC (1998) Bias in discrepant analysis: when two wrongs don’t make a right. J Clin Epidemiol 51(3):219–231. https://doi.org/10.1016/s0895-4356(97)00264-3

38. Blake GM, Fogelman I (2008) How important are BMD accuracy errors for the clinical interpretation of DXA scans? J Bone Miner Res 23(4):457–462. https://doi.org/10.1359/jbmr.071119

Online Resource 1 (Osteoporosis International). Supplementary material for: One DXA scan limits the apparent accuracy of osteoporosis screening tools but explains at most 13% of what FDA-authorized radiograph devices miss

Supplementary Methods S1. Model, fixed-status benchmarks, ROC frontier and corrections

Notation. For a given site, Y is the logarithm of the areal BMD the scanner would report without precision error (the precision-error-free DXA value, not anatomical bone mass) and $W = Y + e$ the logarithm of one measurement, $e \sim N(0, \sigma^2)$, with σ approximated by the root-mean-square CV expressed as a proportion (multiplicative error). $c = \log(\mu - 2.5s)$, with μ and s the young-adult reference mean and SD of BMD; $D = 1[Y \le c]$ (latent DXA-threshold status); $R = 1[W \le c]$ (single-scan DXA label). f(y) is the fitted density of Y (Gaussian mixture with weights w_j, means m_j and SDs v_j) and Φ the standard normal distribution function, so $P(R = 1 \mid Y = y) = \Phi((c - y)/\sigma)$. In this log-scale model, a two-scan reference averages two independent log measurements and has error SD $\sigma/\sqrt{2}$. The young-adult reference mean and standard deviation are 0.858 and 0.120 g/cm² for the femoral neck and 0.942 and 0.122 g/cm² for the total hip, from non-Hispanic White women aged 20–29 years in NHANES III [26]. For L1–L4 we used the Hologic reference for white women, 1.047 and 0.110 g/cm², from a manufacturer communication reported by Melamed et al. [27]. Error structure. In the 1,108 NHANES III femoral-neck repeat pairs (ages ≥20 years), the absolute between-scan difference increased with mean BMD (slope 0.030, $p < 0.001$), whereas the difference of log BMD did not ($p = 0.67$; $p = 0.48$ in the 517 pairs aged ≥50 years). The error was therefore proportional to BMD, as modeled, rather than constant in g/cm² as is often assumed (main-text reference 23). Near the $T = -2.5$ cutoff (mean BMD within 0.08 g/cm²) the RMS-CV was 2.97% (102 pairs; 2.90% in 94 pairs aged ≥50 years), close to the 3.01% NHANES III scenario.

We estimated the distribution of Y in adults aged ≥50 years with valid DXA in NHANES 2005–2006, 2007–2008, 2009–2010, 2013–2014 and the 2017–March 2020 (P_) file (femoral neck n = 12,239; total hip 12,237; lumbar spine 12,144), by site, sex and age group (50–59, 60–69, 70–79, ≥80 years) and pooled. The distributions were survey-weighted Gaussian mixtures fitted by extreme deconvolution with known error variance [28] and averaged over 1–4 components by Bayesian information criterion weights; survey weighting and model averaging are our additions to the published method. For the headline femoral-neck fixed-status benchmarks in women and men, 200 Rao–Wu rescaled bootstrap replicates of primary sampling units [29], each with a refitted distribution, gave percentile 95% confidence intervals (CI); other benchmarks are point

estimates. For other samples, the distribution was shifted to reproduce each sample's DXA-defined prevalence.

Fixed-status benchmarks. For the perfect binary classifier X = D, apparent sensitivity = $P(D = 1 \mid R = 1) = \int_{y \le c} f(y)\, \Phi((c - y)/\sigma)\, dy / \int f(y)\, \Phi((c - y)/\sigma)\, dy$; apparent specificity = $P(D = 0 \mid R = 0) = \int_{y > c} f(y)\, [1 - \Phi((c - y)/\sigma)]\, dy / \int f(y)\, [1 - \Phi((c - y)/\sigma)]\, dy$. κ and PPV follow from the 2 × 2 table of D against R. These values describe the fixed rule X = D, not global maxima over index thresholds. Changing the threshold can increase apparent sensitivity or specificity, and κ need not be maximized at k = c. Separately, under the conditional-independence and monotonicity assumptions below, thresholding the continuous score −Y defines the ROC frontier; its AUC is the probability that Y is lower in a random DXA-positive than in a random DXA-negative person. Integrals were evaluated by composite 12-node Gauss–Legendre quadrature with panel edges at c (and at the index-test threshold k), panels of width σ/2 within ±8σ of each edge and no wider than min(0.01, 2σ) elsewhere. In one cell (femoral neck, women 60–69 years, NHANES III precision) the analytic values agreed with 4×10^6 Monte Carlo draws (largest difference 0.0006; $|z| \le 0.76$ where a standard error was available); 14 unit tests compare every model quantity with Monte Carlo (2×10^6 draws) and with closed forms.

Attainable frontier. For a threshold t on Y, $Se_{app}(t) = P(Y \le t \mid R = 1)$ and $FPR_{app}(t) = P(Y \le t \mid R = 0)$. If an index test's result depends on Y and on noise independent of e, it is conditionally independent of R given Y, so any rule based on it is equivalent to a randomized rule based on Y with the same apparent rates. Because $P(R = 1 \mid Y)$ is monotone in Y, thresholding Y is the most powerful rule at each fixed false-positive rate (Neyman–Pearson lemma), and the test's apparent ROC curve cannot lie above this curve. X = D is the point t = c of this curve, so for a single-site reference no such test exceeds its apparent sensitivity at the same or a lower apparent false-positive rate; for lowest-of-sites references D is not a threshold on a scalar Y and this property need not hold. Table S5 reads the frontier at apparent false-positive rates of 1%, 5% and 10% in each single-site stratum: t was found by root finding on $FPR_{app}(t)$, with a quadrature panel edge placed at t, and at t = c the calculation reproduced the saved benchmarks to within 2.2×10^{-16} (post hoc).

Lowest-of-sites references. True log BMD at the three sites follows a Gaussian copula with the fitted mixtures as marginals and correlations estimated, by sex and age group, from survey-weighted covariances of measured log BMD after removing precision-error variance (and, for the two hip sites, the error covariance of a shared scan). Measurement errors are multivariate normal with correlation 0.30 (95% CI 0.22–0.37) between the hip sites (same scan) and 0.25 between spine and hip (range 0–0.5). D and R are positive if any site is at or below its threshold; each cell used 2,000,000 draws (Monte Carlo SE of sensitivity ≤0.004).

Prevalence matching. All component means m_j are shifted by δ, chosen so that $P(R = 1)$ equals the DXA-defined prevalence of the target sample. Case–control sampling on R

leaves Se_app and Sp_app unchanged but makes the sample prevalence uninformative about the source population. For case–control studies, the fixed-status benchmarks and ROC frontier were therefore evaluated at prevalence 0.872, the highest DXA-defined prevalence among the other audited designs (a generous choice), and corrections used the unshifted US distribution.

Latent-trait correction. The index score is $S = Y + \eta$, $\eta \sim N(0, \tau^2)$ independent of e, and the index result is $X = 1[S \le k]$. Apparent accuracy: $Se_{app}(\tau, k) = \int f(y)\, \Phi((k - y)/\tau)\, \Phi((c - y)/\sigma)\, dy / P(R = 1)$; $Sp_{app}(\tau, k) = \int f(y)\, [1 - \Phi((k - y)/\tau)]\, [1 - \Phi((c - y)/\sigma)]\, dy / P(R = 0)$. True accuracy: $Se(\tau, k) = \int_{y \le c} f(y)\, \Phi((k - y)/\tau)\, dy / P(D = 1)$; $Sp(\tau, k) = \int_{y > c} f(y)\, [1 - \Phi((k - y)/\tau)]\, dy / P(D = 0)$. Given observed (Se_app, Sp_app), (τ, k) are found by root finding; points above the frontier have no exact solution in this model and are flagged as boundary approximations rather than exact inversions. Intervals: delta method on the logit scale using binomial variances of Se_app and Sp_app, or parametric bootstrap. An AUC-based variant solves τ from the apparent AUC.

Conditional-independence (Staquet) correction. With reference sensitivity Se_R, specificity Sp_R and true prevalence p, the correction solves $a = Se\, Se_R\, p + (1 - Sp)(1 - Sp_R)(1 - p)$ and $b = Se\, (1 - Se_R)\, p + (1 - Sp)\, Sp_R\, (1 - p)$ for (Se, 1 − Sp), where $a = P(X = 1, R = 1)$ and $b = P(X = 1, R = 0)$ are observed. With conditional covariances $c_1 = Cov(X, R \mid D = 1)$ and $c_0 = Cov(X, R \mid D = 0)$, the true a exceeds, and the true b falls short of, these expressions by $\Delta = p\, c_1 + (1 - p)\, c_0$. Solving the system therefore shifts the estimates by $\delta Se = (1 - p)\Delta/det$ and $\delta(1 - Sp) = -p\Delta/det$, with $det = p(1 - p)(Se_R + Sp_R - 1) > 0$. Because X and R are conditionally independent given Y and $P(X = 1 \mid Y)$ and $P(R = 1 \mid Y)$ are both non-increasing in Y, $c_1 \ge 0$ and $c_0 \ge 0$ (Harris inequality), so $\Delta \ge 0$ and the correction does not decrease either estimate relative to its true value under these assumptions, and strictly overstates both when $\Delta > 0$.

Simulation. Protocol version 1 was frozen after a 100-replicate pilot (SHA-256 a6e3068a3c833968e1373b25fa781be5bda71f771bd6a91ba18648e2edde2b7a). Amendment 1 added index-error misspecification: t-distributed index error with 3 degrees of freedom, and a mixture in which half the index response depends on continuous BMD and half on binary latent status. Implementation note 1 accelerated the root finding (tabulated starting values, then Newton iteration on the exact forward map; agreement with the original solver $\le 10^{-6}$). Random seeds: SeedSequence(20251025), spawned per scenario. Decision criteria: D1 naive |bias| > 0.02; D2 correction adequate if |bias| ≤ 0.01 and coverage 0.925–0.975 at n ≥ 500 in ≥90% of scenarios; D3 Staquet biased if |bias| > 0.02 in ≥50% of scenarios; D4 latent-trait robust if |bias| ≤ 0.02 under misspecified distribution, reference error or CV; D5 duplicate DXA meets the prespecified naive-estimator criterion if it reduces the root-mean-square error of the naive estimate by ≥25% at equal n; D6 sample size for an interval width ≤0.10; D7 (Amendment 1) latent-trait robust to index-error misspecification if |bias| ≤ 0.02 in all 8 scenarios, otherwise report the latent-trait and Staquet estimates as bounds only if the truth lay between them in all 8.

All three estimators were evaluated with model-known nuisance inputs; the conditional-independence correction was supplied with model reference accuracy and prevalence. The core design crossed two femoral-neck populations (women 60–69, men 70–79 years), DXA CV (1, 2, 3%), true index-test AUC (0.75, 0.85, 0.95), sample size (200, 500, 1,000, 5,000) and design (consecutive, 1:1 case–control) in 144 scenarios; further scenarios varied the BMD distribution, the CV assumed in analysis, the reference-error distribution, site, operating point, duplicate reference scans, AUC input and the index-error model (194 scenarios in total, 2,000 replicates each). The estimand was population true accuracy; a correction was adequate if |bias| ≤ 0.01 and coverage of nominal 95% intervals lay within 0.925–0.975 at n ≥ 500. Sample sizes for a 95% CI no wider than 0.10 were extrapolated from simulated widths assuming width proportional to n^−1/2.

Supplementary extended results

Against one femoral-neck scan at NHANES III precision, the fixed classifier X = D would show sensitivity 0.872 (95% CI 0.862–0.881), specificity 0.990, apparent PPV 0.902 (0.893–0.908) and κ 0.875 (0.865–0.883) in women aged ≥50 years, and sensitivity 0.777 (0.768–0.821; bootstrap mean 0.787) and κ 0.810 (0.804–0.843) in men. In women, 0.9% of the population had latent DXA-threshold osteoporosis but a negative scan and 1.2% a positive scan without it; 99.98% of misclassified women lay within 0.5 T-score units of the threshold (Fig. 2a). Better precision raised these sensitivity benchmarks (women/men: ISCD minimum 0.895/0.818; best case 0.952/0.918; Fig. 2b), as did a two-scan reference (0.912 [0.904–0.918]/0.848 [0.841–0.877]). Across sites, sexes and ages, the NHANES III sensitivity benchmark ranged from 0.695 (total hip, men 60–69 years) to 0.945 (spine, women ≥80 years; Fig. 2d–f) and the ISCD sensitivity benchmark from 0.777 to 0.945. The NHANES III fixed-rule κ benchmark ranged from 0.753 to 0.933, whereas the AUC ceiling was at least 0.992; in men the total hip was below the femoral neck in 3 of 4 age groups. In screening groups the femoral-neck sensitivity benchmark was 0.863 for adults ≥50 years, 0.893 for women ≥65 years and 0.833 for men ≥70 years (Table S12). Defining osteoporosis by the lowest of three sites raised the ≥50-year sensitivity benchmarks from 0.872 to 0.909 in women and from 0.777 to 0.856 in men, partly through higher prevalence. The fixed-rule sensitivity benchmark rose with prevalence, from 0.863 at the US latent prevalence of 5.5% to 0.924 at 33% and 0.970 at 72% (femoral neck, both sexes; Fig. 2c); audited non-case–control studies had a median DXA-defined prevalence of 30.8% and a maximum of 87.2%.

Frontier at fixed false-positive rates (Table S5). The benchmark false-positive rate was 0.01–2.9% at NHANES III precision and 0.01–2.4% at ISCD precision. At an apparent false-positive rate of 1%, frontier sensitivity was 0.833–1.000 at NHANES III precision (0.875–1.000 at ISCD), lowest at the femoral neck in women aged 70–79 years; at 5% it was 0.956–1.000 (0.973–1.000) and at 10% 0.988–1.000 (0.995–1.000). The five strata whose benchmark false-positive rate exceeded 1% (femoral neck, women 70–79 and ≥80 years; total hip, women ≥80; spine, women 70–79 and ≥80) lay below their benchmark at

1%, and the other 19 above it. Moving from NHANES III to ISCD precision raised the 1% frontier by up to 0.042; spine values are identical because both scenarios assume 1.9%.

In the NHANES III subsample (29 DXA-positive participants at the femoral neck, 25 at the total hip), observed femoral-neck agreement between repeat and first scans (sensitivity 0.793, 95% CI 0.603–0.920; κ 0.827, 0.701–0.929) had intervals containing the predictions (0.858; 0.843), and all 12 predicted measures at both sites fell inside the observed CIs with either distribution (Fig. 3). The full-table goodness-of-fit test rejected the examination-based distribution at the femoral neck ($p = 0.009$) but not the subsample-based one ($p = 0.80$). DXA-defined prevalence was 5.6% in the subsample, 11.5% in the survey-weighted examination sample (11.4% unweighted) and 8.0% after unweighted standardization to the subsample's sex–age mix, so predicted counts require a prevalence-matched distribution. The latent-trait correction estimated the repeat scan's femoral-neck error as 0.023 (95% CI 0.000–0.058; scan-pair CV 0.030); at the total hip the observed point lay above the frontier at face value (1 false positive among 490), so the estimate sat at its bound of 0 (upper limit 0.033; scan-pair CV 0.024). The conditional-independence correction returned specificities above 1 in 3 of 4 analyses.

Naive estimates of true sensitivity were biased by −0.032 to +0.005; |bias| exceeded 0.02 in 17 of 144 core scenarios, 16 of them with an index-test AUC of 0.95, concentrating the largest absolute biases in the highest-AUC scenarios tested (Fig. 5a). The conditional-independence correction overestimated true sensitivity in all 144 scenarios (bias +0.014 to +0.112; >0.02 in 134), with coverage of 0.00–0.94 at $n \geq 500$ (Fig. 5b). The latent-trait correction was adequate in all 108 core scenarios with $n \geq 500$ (bias −0.003 to +0.007; coverage 0.940–0.974; specificity bias −0.001 to +0.002) and stayed within 0.009 of the truth under misspecified BMD distribution, CV or reference-error distribution (analyzing at CV 2% when the truth was 3%, or the reverse, gave |bias| 0.0067–0.0085 at $n = 500$ and 1,000, against naive |bias| 0.0076–0.0153 in the same cells), at other sites and operating points, with AUC input, and within 0.004 with heavy-tailed index error. When half the index response depended on binary latent status rather than continuous BMD, the latent-trait bias was −0.032 to −0.020 (beyond 0.02 in 3 of 4), removing 24–31% of the naive bias, so the pre-specified robustness criterion failed; the conditional-independence correction then overestimated by 0.024–0.036, and the truth lay between the two corrections in all 4 scenarios. The fallback of treating the corrections as bounds also failed: in one heavy-tailed scenario the latent-trait estimate exceeded the truth by 0.002 (7 of 8 misspecified scenarios bracketed; Fig. 5c). In the original duplicate-reference scenarios (women 60–69 years, index AUC 0.85, CV 2–3%, $n = 500$ or 1,000), a two-scan log-mean reference changed the root-mean-square error of naive sensitivity by −2.6% to +0.7% at equal sample size and increased it by 36–48% at an equal number of scans. A 95% CI no wider than 0.10 for corrected sensitivity would need an estimated 4,199–5,132 consecutive women aged 60–69 years, 10,909–12,116 consecutive men aged 70–79 years, or 621–806 participants in 1:1 case–control designs (Table S11).

Precision assessment. A 30-pair precision study estimates the RMS-CV with relative standard error of about $1/\sqrt{60} = 0.13$, i.e. 0.39 CV points at a CV of 3%. At about 0.8 points of latent-trait sensitivity bias per CV point, the resulting uncertainty is about 0.3 points; added in quadrature to the latent-trait sampling SD at n = 1,000 in the same scenarios (5.7–5.8 points), it increases that SD by about 0.15%.

Supplementary Methods S2. Literature search and audit

Search. PubMed was searched on 25 October 2026 through the E-utilities interface (title and abstract fields; publication dates from 1 January 2019) with six queries: artificial-intelligence imaging (n = 351); opportunistic CT (n = 369); radiographic indices and texture (n = 212); REMS and quantitative ultrasound (n = 83); biomarkers (n = 339); systematic reviews and meta-analyses (n = 84). The union contained 1,192 unique PMIDs, of which 1,191 records were retrieved. The queries are listed at the end of this section.

Eligibility. Primary studies in adults were eligible if the index test was not the DXA measurement itself (imaging, ultrasound or laboratory tests, alone or combined with clinical variables), the reference standard was central DXA with osteoporosis defined as $T \leq -2.5$ or by the WHO criteria at one or more sites (including the lowest of several central sites), and sensitivity, specificity, AUC, overall accuracy, predictive values or a 2 × 2 table for osteoporosis versus no osteoporosis (or versus normal BMD) was reported in the abstract or full text. Reviews, animal and phantom studies, studies of children only, studies with fracture or other clinical outcomes as the target, index tests derived from the same DXA scan, other reference standards or thresholds (forearm DXA, CT or QCT measurements, Z-scores, combined osteopenia–osteoporosis targets or definitions that add fracture criteria) and studies without classification accuracy were excluded. Systematic reviews and meta-analyses of eligible tests were collected separately.

Screening. The 1,191 records yielded 387 provisional inclusions and 133 candidate systematic reviews.

Extraction. Open-access full texts (Europe PMC; 198 of the 387) were searched for reporting items: threshold, reference-site definition, DXA precision and repeat DXA. Each of the 133 candidate reviews was checked against its abstract, where 11 meta-analyses reported pooled sensitivity and specificity for osteoporosis. Five were not corrected: one reported three threshold-specific pairs rather than one pooled estimate, one pooled mainly studies that used the mandibular cortical index rather than DXA as the reference, one did not require a DXA reference, one did not state its reference standard in the abstract and its full text is not openly available, and one pooled models applied to DXA images, whose errors are shared with the reference. The remaining 6 were corrected; their pooled values and reference standards were checked against the abstract or full text (Table S2).

Eligibility and data rules. Accuracy values were taken from the abstract, for the study's own index test, from the first evaluation set (including internal held-out sets), with sensitivity and specificity at the same operating point; counts could also come from the

full text and were kept only if they reproduced the reported sensitivity or specificity within rounding (they did not in 7 studies). DXA precision counted as mentioned when a study gave a precision value (CV, least significant change or RMS-SD), described precision monitoring, named precision error as a limitation or assumed no measurement error; general statements about DXA or its calibration did not count. Repeat DXA counted when scans were repeated within a short interval, not when follow-up scans measured change. Of the 387, 228 were eligible; 103 had a reference other than central DXA or a target other than DXA-defined osteoporosis, 25 reported no accuracy value for osteoporosis, and 31 could not be classified from the abstract and, for 5, the full text. Thirty eligible studies contribute no sensitivity, specificity or AUC: 24 report accuracy only in the full text and 6 report only other measures.

Comparison with fixed-status benchmarks and frontiers. DXA-defined prevalence was computed from the reported numbers of participants with and without DXA-defined osteoporosis in the set that gave the accuracy values. Specificity and AUC depend on who forms the comparison group, so they were compared only when it comprised all participants without osteoporosis; when it was restricted to normal BMD (25 studies) or osteopenia (3) or was unclear (2), the reported values were kept but not compared; such studies were treated as case–control when they sampled by BMD status and otherwise used their whole-sample prevalence. Fixed-status benchmarks were matched using the highest benchmark over the three distribution shapes (women, men, both sexes) and, when the reference site was unclear or several sites were used, over the candidate sites; case–control studies were evaluated at prevalence 0.872. Reported sensitivity and specificity were compared with their fixed-status benchmarks by one-sided exact binomial tests. These are descriptive comparisons with X = D: exceeding one benchmark is expected when an index threshold trades sensitivity for specificity, so they do not test compatibility with the ROC frontier. Operating points with specificity were compared with the femoral-neck frontier for both sexes at NHANES III precision and with the generous frontier (maximum over candidate sites and shapes at best-case precision); a point counted as significantly above only if the exact 95% lower limit of its sensitivity exceeded the generous frontier at the exact 95% upper limit of its false-positive rate (1 of 29 above the NHANES III frontier at face value, none above the generous frontier). Reported AUCs (n = 132) were compared with the generous AUC ceiling at face value, using the study's prevalence when known (63 studies) and 0.872 otherwise; one AUC (1.000, serum apelin-13 in a case–control study, PMID 42447047) exceeded its ceiling at best-case (0.9995) and NHANES III precision (0.997), and the smallest NHANES III AUC ceiling was 0.992. Corrections used ISCD precision and the sex-pooled distribution at the reference site (femoral neck when the definition mentioned the hip or femur or was unclear, otherwise the spine), matched to the study's prevalence; sensitivity analyses varied site, precision and sex, and shifted case–control studies to the median prevalence of the other designs (0.308).

Audit results. Median reported AUC was 0.84 (interquartile range 0.77–0.91; n = 152), sensitivity 0.824 (0.714–0.907; n = 106) and specificity 0.829 (0.730–0.905; n = 103). DXA-defined prevalence was computable for 113 studies (median 0.337; 0.308 outside case–control designs, maximum 0.872). Among 38 studies with sensitivity and prevalence (four case–control), the median matched sensitivity benchmark at NHANES III precision was 0.949. Seven sensitivity and one specificity point estimates exceeded their benchmarks at NHANES III precision (five and one at ISCD, one and none at best-case precision); exceeding one benchmark is expected when an index threshold trades sensitivity for specificity, and none did so significantly. 28 of 29 latent-trait fits were interior (the 16-of-16 point lay on the boundary); median sensitivity increased by 0.004 (interquartile range 0.003–0.012; maximum 0.032), compared with 0.039 (maximum 0.079) under conditional-independence correction, which returned sensitivities above 1 for four studies.

Queries.

Artificial-intelligence imaging: (osteoporosis[tiab] OR osteoporotic[tiab] OR "low bone mass"[tiab] OR "low bone mineral density"[tiab]) AND ("deep learning"[tiab] OR "convolutional neural"[tiab] OR "artificial intelligence"[tiab] OR "machine learning"[tiab] OR "neural network"[tiab] OR radiomic*[tiab]) AND (radiograph*[tiab] OR "X-ray"[tiab] OR xray[tiab] OR "computed tomography"[tiab] OR CT[tiab] OR panoramic[tiab] OR "magnetic resonance"[tiab] OR MRI[tiab] OR image*[tiab] OR imaging[tiab]) AND (DXA[tiab] OR DEXA[tiab] OR "dual-energy"[tiab] OR "dual energy"[tiab] OR densitometry[tiab] OR "T-score"[tiab] OR "T score"[tiab] OR "bone mineral density"[tiab] OR BMD[tiab]) AND (sensitivity[tiab] OR specificity[tiab] OR AUC[tiab] OR "area under"[tiab] OR "diagnostic accuracy"[tiab] OR "diagnostic performance"[tiab] OR accuracy[tiab] OR "receiver operating"[tiab] OR ROC[tiab]) AND ("2019/01/01"[dp] : "3000"[dp]) NOT (animals[mh] NOT humans[mh])

Opportunistic CT: (osteoporosis[tiab] OR osteoporotic[tiab] OR "low bone mass"[tiab] OR "low bone mineral density"[tiab]) AND (opportunistic[tiab] OR "Hounsfield unit"[tiab] OR "Hounsfield units"[tiab] OR HU[tiab] OR attenuation[tiab]) AND ("computed tomography"[tiab] OR CT[tiab]) AND (DXA[tiab] OR DEXA[tiab] OR "dual-energy"[tiab] OR "dual energy"[tiab] OR densitometry[tiab] OR "T-score"[tiab] OR "T score"[tiab] OR "bone mineral density"[tiab] OR BMD[tiab]) AND (sensitivity[tiab] OR specificity[tiab] OR AUC[tiab] OR "area under"[tiab] OR "diagnostic accuracy"[tiab] OR "diagnostic performance"[tiab] OR accuracy[tiab] OR "receiver operating"[tiab] OR ROC[tiab]) AND ("2019/01/01"[dp] : "3000"[dp]) NOT (animals[mh] NOT humans[mh])

Radiographic indices and texture: (osteoporosis[tiab] OR osteoporotic[tiab] OR "low bone mass"[tiab] OR "low bone mineral density"[tiab]) AND (radiograph*[tiab] OR "X-ray image"[tiab] OR "X-ray images"[tiab] OR "plain film"[tiab] OR panoramic[tiab] OR "cortical thickness"[tiab] OR "cortical width"[tiab] OR "mandibular cortical"[tiab] OR texture[tiab] OR fractal[tiab] OR "Singh index"[tiab]) AND (DXA[tiab] OR DEXA[tiab]

OR "dual-energy"[tiab] OR "dual energy"[tiab] OR densitometry[tiab] OR "T-score"[tiab] OR "T score"[tiab] OR "bone mineral density"[tiab] OR BMD[tiab]) AND (sensitivity[tiab] OR specificity[tiab] OR AUC[tiab] OR "area under"[tiab] OR "diagnostic accuracy"[tiab] OR "diagnostic performance"[tiab] OR accuracy[tiab] OR "receiver operating"[tiab] OR ROC[tiab]) AND ("2019/01/01"[dp] : "3000"[dp]) NOT (animals[mh] NOT humans[mh]) NOT ("deep learning"[tiab] OR "convolutional neural"[tiab] OR "artificial intelligence"[tiab] OR "machine learning"[tiab])
REMS and quantitative ultrasound: (osteoporosis[tiab] OR osteoporotic[tiab] OR "low bone mass"[tiab] OR "low bone mineral density"[tiab]) AND ("radiofrequency echographic"[tiab] OR REMS[tiab] OR "quantitative ultrasound"[tiab] OR QUS[tiab] OR "calcaneal ultrasound"[tiab] OR "ultrasound"[ti]) AND (DXA[tiab] OR DEXA[tiab] OR "dual-energy"[tiab] OR "dual energy"[tiab] OR densitometry[tiab] OR "T-score"[tiab] OR "T score"[tiab] OR "bone mineral density"[tiab] OR BMD[tiab]) AND (sensitivity[tiab] OR specificity[tiab] OR AUC[tiab] OR "area under"[tiab] OR "diagnostic accuracy"[tiab] OR "diagnostic performance"[tiab] OR accuracy[tiab] OR "receiver operating"[tiab] OR ROC[tiab]) AND ("2019/01/01"[dp] : "3000"[dp]) NOT (animals[mh] NOT humans[mh])
Biomarkers: (osteoporosis[tiab] OR osteoporotic[tiab] OR "low bone mass"[tiab] OR "low bone mineral density"[tiab]) AND (biomarker*[tiab] OR "bone turnover"[tiab] OR CTX[tiab] OR P1NP[tiab] OR PINP[tiab] OR osteocalcin[tiab] OR sclerostin[tiab] OR microRNA*[tiab] OR miRNA*[tiab] OR metabolom*[tiab] OR proteom*[tiab] OR "serum"[tiab] OR "plasma"[tiab] OR urinary[tiab]) AND (DXA[tiab] OR DEXA[tiab] OR "dual-energy"[tiab] OR "dual energy"[tiab] OR densitometry[tiab] OR "T-score"[tiab] OR "T score"[tiab] OR "bone mineral density"[tiab] OR BMD[tiab]) AND ("diagnostic"[tiab] OR "diagnosis"[tiab] OR screening[tiab] OR identif*[tiab] OR discriminat*[tiab]) AND (sensitivity[tiab] OR specificity[tiab] OR AUC[tiab] OR "area under"[tiab] OR "receiver operating"[tiab] OR ROC[tiab]) AND ("2019/01/01"[dp] : "3000"[dp]) NOT (animals[mh] NOT humans[mh])
Systematic reviews and meta-analyses: (osteoporosis[tiab] OR osteoporotic[tiab] OR "low bone mass"[tiab] OR "low bone mineral density"[tiab]) AND (DXA[tiab] OR DEXA[tiab] OR "dual-energy"[tiab] OR "dual energy"[tiab] OR densitometry[tiab] OR "T-score"[tiab] OR "T score"[tiab] OR "bone mineral density"[tiab] OR BMD[tiab]) AND (sensitivity[tiab] OR specificity[tiab] OR AUC[tiab] OR "area under"[tiab] OR "diagnostic accuracy"[tiab] OR "diagnostic performance"[tiab] OR pooled[tiab]) AND ("deep learning"[tiab] OR "artificial intelligence"[tiab] OR "machine learning"[tiab] OR opportunistic[tiab] OR "Hounsfield"[tiab] OR radiomic*[tiab] OR radiograph*[tiab] OR "quantitative ultrasound"[tiab] OR REMS[tiab] OR "radiofrequency echographic"[tiab] OR biomarker*[tiab] OR "bone turnover"[tiab] OR panoramic[tiab]) AND (meta-analysis[pt] OR systematic review[pt] OR "meta-analysis"[tiab] OR "systematic review"[tiab]) AND ("2019/01/01"[dp] : "3000"[dp]) NOT (animals[mh] NOT humans[mh])

Table S1. Studies with reported sensitivity and DXA-defined prevalence: prevalence-matched fixed-status benchmarks for sensitivity and specificity (maximum over candidate sites and shapes), frontier check and corrections (primary setting: ISCD precision)

PMID	Index test	Case–control sampling	Reference definition	n	DXA prevalence	Se	Sp	X=D Se benchmark NHANES III/ISCD/best (exact p, ISCD)	X=D Sp benchmark NHANES III/ISCD/best (exact p, ISCD)	Above frontier	LT-corrected Se (95% CI)	Staquet-corrected Se	Comparison group
42877638	AI imaging	no	LS (L1–L4)	3007	0.067	0.910	0.827	0.918/0.918/0.959 (0.714)	0.995/0.995/0.997 (1.000)	no / no	0.916 (0.867–0.948)	0.976	all
39934075	Opportunistic CT	no	LS (levels not stated)	159	0.082	0.923	0.986	0.921/0.921/0.960 (0.723)	0.994/0.994/0.997 (0.945)	yes / no	0.955 (0.293–0.999)	1.002 (Se > 1)	all
38846273	Opportunistic CT	no	total femur	291	0.103	0.633	0.981	0.908/0.932/0.971 (1.000)	0.991/0.993/0.997 (0.992)	no / no	0.652 (0.458–0.806)	0.701	all
35836165	Biomarker	no	TH, FN or LS (any)	252	0.111	0.880	0.630	0.925/0.933/0.971 (0.886)	0.992/0.993/0.997 (1.000)	no / no	0.886 (0.703–0.962)	0.936	all
32218298	REMS or QUS	no	LS (L1–L4) or hip (either)	786	0.123	0.144	0.961	0.927/0.935/0.972 (1.000)	0.991/0.992/0.996 (1.000)	no / no	0.146 (0.088–0.232)	0.155	all
39934236	Opportunistic CT	no	not stated	219	0.142	0.840	0.900	0.929/0.937/0.973 (0.988)	0.990/0.991/0.996 (1.000)	no / no	0.854 (0.671–0.944)	0.915	all
31731334	Opportunistic CT	no	LS (L1–L4 mean)	111	0.162	0.900	–	0.932/0.932/0.965 (0.880)	0.989/0.989/0.994 (–)				all
41450962	Radiographic index or texture	no	lowest of FN and LS	241	0.162	0.660	0.940	0.932/0.932/0.965 (1.000)	0.989/0.989/0.994 (1.000)	no / no	0.673 (0.506–0.805)	0.718	all
38917853	Opportunistic CT	no	LS (L1–L4) or hip (lower)	470	0.191	0.544	–	0.935/0.943/0.975 (1.000)	0.987/0.988/0.994 (–)				osteopenia
33661911	Opportunistic CT	no	FN (lowest)	150	0.200	0.800	0.958	0.902/0.919/0.962 (0.991)	0.980/0.983/0.991 (0.982)	no / no	0.820 (0.619–0.927)	0.868	all
36833090	Opportunistic CT	no	lowest of L1–L4	105	0.219	0.957	0.756	0.939/0.939/0.968 (0.583)	0.985/0.985/0.992 (1.000)	no / no	0.962 (0.734–0.996)	1.004 (Se > 1)	all
39443535	Opportunistic CT	no	not stated (LS and hip scanned)	241	0.220	0.654	0.709	0.939/0.946/0.976 (1.000)	0.985/0.986/0.994 (1.000)	no / no	0.658 (0.520–0.774)	0.685	all
34982780	Opportunistic CT	no	FN (lowest)	430	0.223	0.948	0.850	0.906/0.923/0.964 (0.239)	0.977/0.981/0.990 (1.000)	no / no	0.960 (0.885–0.987)	1.016 (Se > 1)	all
42274936	REMS or QUS	no	FN for Se (sites analysed separately)	148	0.236	0.460	–	0.940/0.947/0.977 (1.000)	0.984/0.985/0.993 (–)				all
30984666	Radiographic index or texture	no	spine (levels not stated)	62	0.258	1.000	0.391	0.943/0.943/0.970 (0.389)	0.982/0.982/0.990 (1.000)	on / no	1.000 (–)‡	1.024 (Se > 1)	all
37093523	REMS or QUS	no	hip (region not stated)	333	0.267	0.944	0.590	0.933/0.950/0.978 (0.715)	0.978/0.983/0.992 (1.000)	no / no	0.948 (0.875–0.980)	0.986	all
33545629	Opportunistic CT	no	LS and hip (rule not stated)	246	0.268	0.950	–	0.944/0.950/0.978 (0.581)	0.981/0.983/0.992 (–)				all
40660211	Biomarker	no	not stated (LS, FN and hip scanned)	126	0.270	0.794	0.783	0.944/0.944/0.971 (1.000)	0.981/0.981/0.990 (1.000)	no / no	0.801 (0.629–0.905)	0.839	all

PMID	Index test	Case–control sampling	Reference definition	n	DXA prevalence	Se	Sp	X=D Se benchmark NHANES III/ISCD/best (exact p, ISCD)	X=D Sp benchmark NHANES III/ISCD/best (exact p, ISCD)	Above frontier	LT-corrected Se (95% CI)	Staquet-corrected Se	Comparison group
37274169	Radiographic index or texture	no	LS (L1–L4)	104†	0.308	0.936	–	0.948/0.948/0.973 (–)	0.978/0.978/0.988 (–)				normal BMD
41549370	Opportunistic CT	no	lowest of LS and FN	180	0.333	0.600	0.783	0.950/0.950/0.974 (1.000)	0.977/0.977/0.987 (1.000)	no / no	0.604 (0.474–0.720)	0.626	all
37644799	Opportunistic CT	no	LS	210	0.362	0.908	0.881	0.952/0.952/0.975 (0.972)	0.975/0.975/0.986 (1.000)	no / no	0.916 (0.821–0.963)	0.948	all
37495426	AI imaging	no	LS	145	0.379	0.782	0.822	0.954/0.954/0.976 (1.000)	0.973/0.973/0.986 (1.000)	no / no	0.786 (0.656–0.876)	0.812	all
42724648	AI imaging	no	LS, FN or TH (any)	52	0.385	0.900	0.570	0.954/0.959/0.982 (0.954)	0.973/0.976/0.989 (1.000)	no / no	0.904 (0.672–0.977)	0.929	all
42832076	Radiographic index or texture	no	lowest of LS and FN	162	0.389	0.825	0.576	0.955/0.955/0.976 (1.000)	0.973/0.973/0.985 (1.000)	no / no	0.828 (0.713–0.904)	0.850	all
33017775	Opportunistic CT	no	LS (implied)	53	0.434	0.960	0.930	0.958/0.958/0.978 (0.753)	0.969/0.969/0.984 (0.936)	no / no	0.975 (0.477–0.999)	1.000	all
42016465	Biomarker	no	LS (L1–L4)	324	0.435	0.702	0.738	0.959/0.959/0.978 (1.000)	0.969/0.969/0.984 (1.000)	no / no	0.704 (0.623–0.774)	0.722	all
42482899	Opportunistic CT	no	LS or FN (any)	262	0.447	0.513	0.876	0.960/0.960/0.979 (1.000)	0.968/0.968/0.983 (1.000)	no / no	0.516 (0.425–0.606)	0.534	all
35433140	Opportunistic CT	no	LS and hip (rule not stated)	30	0.467	0.928	0.812	0.961/0.965/0.984 (0.914)	0.967/0.970/0.986 (0.999)	no / no	0.936 (0.587–0.993)	0.967	all
36517540	Opportunistic CT	no	L1	239	0.469	0.714	0.693	0.961/0.961/0.980 (1.000)	0.966/0.966/0.982 (1.000)	no / no	0.716 (0.625–0.792)	0.731	all
36896411	Radiographic index or texture	no	not stated (femur and spine scanned)	42	0.476	0.900	0.909	0.962/0.965/0.985 (0.969)	0.966/0.969/0.986 (0.971)	no / no	0.912 (0.643–0.983)	0.942	all
42881638	AI imaging	no	lowest of LS (L1–L4) and FN	2336	0.512	0.697	0.846	0.965/0.965/0.981 (1.000)	0.963/0.963/0.980 (1.000)	no / no	0.700 (0.673–0.726)	0.723	all
34620976	AI imaging	no	proximal femur (same side)	101	0.554	0.911	0.689	0.960/0.970/0.987 (0.994)	0.950/0.963/0.983 (1.000)	no / no	0.915 (0.803–0.966)	0.937	all
42426752	Opportunistic CT	no	lowest of LS, TH and FN	145	0.614	0.809	0.857	0.972/0.974/0.988 (1.000)	0.954/0.957/0.981 (0.999)	no / no	0.812 (0.715–0.882)	0.834	all
37845656	REMS or QUS	no	LS (L2–L4) or FN (either)	67	0.716	0.104	0.947	0.979/0.979/0.989 (1.000)	0.943/0.943/0.970 (0.702)	no / no	0.104 (0.044–0.227)	0.105	all
37851645	Biomarker	yes	FN and LS (rule not stated)	126	0.516	0.831	–	0.990/0.990/0.994 (1.000)	0.918/0.918/0.958 (–)				normal BMD
39263180	Biomarker	yes	FN or LS (either)	201	0.502	0.832	–	0.990/0.990/0.994 (1.000)	0.918/0.918/0.958 (–)				normal BMD
39434260	Biomarker	yes	not stated	142	0.500	0.840	–	0.990/0.990/0.995 (1.000)	0.918/0.918/0.965 (–)				normal BMD
41645207	Opportunistic CT	yes	LS (L1–L4)	100	0.500	0.860	–	0.990/0.990/0.994 (1.000)	0.918/0.918/0.958 (–)				normal BMD

Fixed-status benchmarks use the maximum over candidate reference sites and over the women's, men's and pooled distributions; case–control studies are evaluated at prevalence 0.872, the highest prevalence among the other designs. Exact p: descriptive one-sided binomial comparison of the reported Se or Sp with its ISCD fixed-status benchmark; exceeding one benchmark can reflect a threshold trade-off, so this is not a frontier-compatibility test; –, Sp not reported or not compared. Above frontier: at face value, femoral-neck frontier at NHANES III precision / generous frontier; on, a reported sensitivity of 1.000 on a frontier of 1.000. †n and prevalence from the whole sample; sensitivity from a 12-image test set whose osteoporosis count is not reported, so no exact p. ‡Interval undefined at a reported sensitivity of 1.000 (16 of 16). Comparison group: all, all participants without osteoporosis; normal BMD or osteopenia, a restricted group against which Sp is not compared. LT, latent-trait correction; Se > 1, the conditional-independence (Staquet) correction returned a sensitivity above 1. AI, artificial intelligence; BMD, bone mineral density; CT, computed tomography; FN, femoral neck; LS, lumbar spine; QUS, quantitative ultrasound; REMS, radiofrequency echographic multispectrometry; Se, sensitivity; Sp, specificity; TH, total hip.

Table S2. Pooled sensitivity of meta-analyses corrected by the latent-trait (LT) and conditional-independence (Staquet) methods, ISCD precision

PMID	Index test	Reference standard	Pooled Se/Sp	FN 10%: LT/Staquet Se	FN 37%: LT/Staquet Se	FN 60%: LT/Staquet Se	LS 10%: LT/Staquet Se	LS 37%: LT/Staquet Se	LS 60%: LT/Staquet Se
34530357 [3][a]	thoracic/lumbar CT HU thresholds	DXA T ≤ −2.5, or moderate–severe compression fracture in 3 of 10 studies (full text)	0.830/0.740	0.838/0.894	0.836/0.867	0.833/0.852	0.835/0.878	0.834/0.859	0.832/0.848
38248083 [4]	deep learning plain X-rays	DXA, T ≤ −2.5 (full text)	0.810/0.870	0.822/0.886	0.818/0.854	0.814/0.837	0.817/0.867	0.816/0.844	0.813/0.831
39633356 [5]	deep learning (CNN) osteoporosis prediction	DXA in 9 of 10 studies, CT in 1; threshold not stated (full text)	0.860/0.890	0.874/0.944	0.871/0.908	0.865/0.889	0.869/0.923	0.867/0.898	0.864/0.884
42677089 [S1][b]	circRNAs	DXA, WHO criteria or clinical description (full text)	0.870/0.790	0.879/0.944	0.877/0.912	0.874/0.896	0.876/0.925	0.875/0.903	0.873/0.891
42810321 [S2]	MRI VBQ	DXA (abstract)	0.670/0.750	0.676/0.717	0.674/0.697	0.672/0.686	0.674/0.705	0.672/0.691	0.671/0.683
31596133 [S3]	panoramic mandibular cortical index (MCI)	DXA (abstract)	0.350/0.880	0.354/0.376	0.352/0.365	0.351/0.359	0.352/0.369	0.351/0.362	0.350/0.357

circRNA, circular RNA; CNN, convolutional neural network; CT, computed tomography; FN, femoral neck; HU, Hounsfield units; LS, lumbar spine; MCI, mandibular cortical index; MRI, magnetic resonance imaging; VBQ, vertebral bone quality; WHO, World Health Organization; percentages are assumed DXA-defined prevalence. Pooled values as reported. Sources: references 3–5 (main text) and [S1]–[S3]. [a]Unadjusted pooled estimates; the review also reports bias-adjusted sensitivity and specificity of 0.73 and 0.71. [b]Predominantly case–control studies. Excluded from correction: PMID 33722727, reported three threshold-specific pooled pairs rather than one pooled estimate; PMID 40075328, pooled mainly studies with the mandibular cortical index, not DXA, as reference (4 of 7); PMID 40010536, its inclusion criteria did not require a DXA reference; PMID 37773681, reference standard not stated in the abstract, and the full text is not openly available; PMID 37430259, index tests are models applied to DXA images, whose errors are shared with the reference.

Table S3. Precision scenarios

Scenario	FN CV %	TH CV %	LS CV %	Source
NHANES III (1988–1994)	3.01	2.40	1.90	NHANES III (1988–1994) second-examination repeat hip scans, adults ≥50 y (survey quality-control report: femoral neck 3.2% in 535 duplicate scans aged 20–91 y; main-text reference 30). Spine assumed from ISCD (not scanned in NHANES III). Not an estimate of contemporary scanner performance.
ISCD minimum acceptable technologist precision	2.50	1.80	1.90	2023 ISCD Official Adult Positions (main-text reference 1); the worst precision acceptable before retraining

Scenario	FN CV %	TH CV %	LS CV %	Source
Assumed best case	1.20	0.80	1.00	Assumption, below the NHANES III and ISCD values at every site; published good practice is about 1.5% at the femoral neck and 1.0–1.3% at the spine (main-text references 19, 23)

CV, coefficient of variation; FN, femoral neck; LS, lumbar spine; TH, total hip. These are the original modeling scenarios. The observed bone analysis uses a separate companion-study extraction: 1,108 adult pairs, 517 participants aged ≥50 years, overall RMS-CV 2.8404%, and an arithmetic mean on the original BMD scale. Its precision estimate is not interchangeable with the 3.01% age-specific NHANES III scenario.

Supplementary Methods S3. Reference design and observed allocation comparisons

This post hoc extension compares reference-allocation policies, estimator-specific performance and observed label preservation. Analyses of the same survey data and nested age groups are descriptive companion analyses, not independent replications. Supplementary Methods S1 and S2 describe the benchmark, correction-simulation and literature-audit methods.

Reference rules and estimator specific targets

For bone, the six compared rules were a single measurement; a mean of two for everyone; a mean of two only inside the fixed ±2 error-unit window; positive-triggered two-of-two; positive-triggered up-to-three; and sequential two-of-three majority. The two-of-two rule retains a negative first result and requires a positive second result after a positive first. Up-to-three also retains a negative first result; after a positive first result, it accepts a positive second result or, if the second is negative, uses a third result for confirmation. Sequential majority measures everyone twice and requests a third measurement only when the first two dichotomous results disagree. No window was optimized against these reported outcomes.

The fixed-index bone calculations used $c = \log(0.558\ g/cm^2)$, derived from the reference mean and SD above, with the generating index cutoff $k = c$. The observed-pair allocation comparison below used its own cutoff of 0.5576273005137747 g/cm² and arithmetic-BMD averaging. These analyses are not numerical replications of each other.

These rules define reference labels for an accuracy study. They are mathematical comparisons rather than complete clinical diagnostic pathways; the bias findings do not establish whether clinical confirmation should be performed.

For the fasting plasma glucose (FPG) model contrast, we reused the saved latent Gaussian mixture and reference-error Gaussian mixture on the reversed natural-log scale, $Y = -\log(FPG)$. True positive status was $Y \le c$, where c corresponded to 125.95 mg/dL. The index followed the same Gaussian family as the bone analysis: $S = Y + \eta$, with η normally distributed and independent of reference errors conditional on Y, and $X = 1[S \le k]$. Its error scale was calibrated once to each true AUC target, 0.85 and 0.95, within the frozen FPG distribution; that scale and $k = c$ were then held fixed across reference policies. Each repeat error was independently drawn from the saved mixture. Universal averaging used the full independent two-error convolution, rather than replacing the mixture by a Gaussian with reduced RMS. The eight policies comprised a single measurement, selective averaging within one, two or three RMS units, universal averaging, positive-triggered two-of-two, positive-triggered up-to-three and sequential majority. These were fixed index scenarios, not trained classifiers or observed predictions.

For the naive-estimator comparison, the expected reference-measurement budgets were B = 500, 1,000 and 5,000. Each policy evaluated a fixed number of independent participants $n = \text{floor}(B/E[N])$, where E[N] was its expected number of reference measurements per participant; index-test costs were excluded. This construction limits expected expenditure and does not impose a hard stopping cap on the realized measurement count. The sensitivity denominator was $K \sim \text{Binomial}(n, P(R = 1))$; specificity used $P(R = 0)$. Each endpoint was evaluated conditional

on its own $K > 0$ event. With apparent accuracy a and fixed true accuracy t, the exact conditional MSE was $(a - t)^2 + a(1 - a)E[1/K \mid K > 0]$, with the reciprocal expectation evaluated by a finite binomial sum. All latent-distribution and error parameters were held fixed. Because this MSE adds the binomial term $a(1 - a)E[1/K \mid K > 0]$ to the squared bias, a rule that moves apparent accuracy towards 1 lowers that term whatever it does to the bias; signed bias is therefore reported with every RMSE, rules are not ranked by RMSE alone, and at large budgets, where MSE approaches the squared bias, |bias| is the relevant criterion. These finite-n naive losses are separate from the corrected-estimator comparison, which reports local asymptotic covariance per expected reference budget under known nuisance inputs.

For the corrected comparison, let a and b denote apparent sensitivity and specificity, p the reference-positive probability, and $\theta = (\tau, k)$. Let J be the Jacobian of (a, b) with respect to θ, and H the Jacobian of true sensitivity and specificity. At an identifiable interior point, $G = H J^{-1}$. The per-participant covariance coefficient is $V = G \, \mathrm{diag}[a(1-a)/p, b(1-b)/(1-p)] \, G^T$. Multiplication by expected measurements per participant gives the covariance coefficient per expected measurement budget. Ratios in Fig. S2 divide these coefficients by their single-reference values. This calculation holds the calibration inputs fixed and propagates sampling variation in the two binary accuracy ratios. It is not a likelihood analysis of the continuous measurements.

For misspecification, the existing mixed index generated the apparent sensitivity and specificity, and the Gaussian latent-trait inversion returned pseudo-true parameters under each reference rule. Bias was the resulting population accuracy minus the generating true accuracy. The original 50% latent-trait and 50% disease-class component, reference CV of 3%, and generating parameters were retained. No finite-sample variance, coverage or budget crossover was computed for this secondary comparison.

Paired bone data

The observed femoral-neck policy comparison used the separately screened companion-study extraction of NHANES III repeat examinations: acceptable scans at both visits, BMD between 0.2 and 2.0 g/cm² and age ≥20 years yielded 1,108 pairs, including 517 participants aged ≥50. This was a different analytic extraction of the same survey from the 515-participant sample used in the original agreement analysis; the extracts were not pooled and do not constitute independent replication. Observed positivity was BMD ≤0.5576273005137747 g/cm². Within a window of ±2 × (RMS-CV/100) × cutoff around the first BMD, the selective rule used the arithmetic mean of the two BMD measurements; outside it, the first label was retained. The saved overall RMS-CV of 2.8404007234788753% came from these same 1,108 pairs and was used in both age scopes. Thus this observed arithmetic-BMD rule was distinct from the log-BMD averaging used in the model calculations. Counts were unweighted, and the nested age scopes shared participants.

Additional results and boundaries

Bone budget and corrected comparisons. At 5,000 expected scans in the CV 3%, AUC 0.95 scenario (true sensitivity 0.7273), naive sensitivity RMSE was 0.03790 with one scan and 0.02895 with selective averaging (bias −2.97 and −1.52 points); a positive-triggered up-to-three rule reached 0.02525 only because opposing errors cancelled (bias +0.26 points here, negative in

the glucose model). The selective window lowered sensitivity RMSE in five of 12 scenario–budget combinations, by 0.6–23.6%, and increased it in seven, by 0.7–4.5%; universal duplication increased it in 11 of 12. Single measurement minimized sensitivity RMSE in seven combinations and specificity RMSE in all 12. Under correctly specified latent-trait correction, the glucose models reversed the bone sensitivity result, with decreases in variance per scan of 3.5% and 11.9%, but not the specificity result; universal averaging had higher variance per budget than either alternative for both endpoints in all four scenarios. Under the separate mixed-index mechanism, corrected sensitivity bias was −3.136 percentage points with one scan, −2.269 with universal averaging and −2.270 with selective averaging (Fig. S2c): the selective rule, at 1.105 scans per participant, removed 27.6% of the bias and left 2.27 points.

The FPG model illustrated why the reference rule with the lowest evaluation error depends on the endpoint. At true index AUC 0.95 and an expected budget of 5,000 reference measurements, conditional sensitivity RMSE was 0.079931 with a single reference, 0.040702 with the two-RMS window and 0.048532 with universal averaging. Positive-triggered up-to-three had the lowest sensitivity RMSE among the eight policies, 0.028683, but its bias was negative here and positive in all four bone scenarios (+0.26 to +0.59 points), and its low RMSE partly reflects the smaller binomial variance of an apparent sensitivity nearer 1. In bone, sequential majority-of-three behaved like universal averaging (sensitivity bias −0.32 to −1.37 points for 2.02–2.03 scans per participant). Single-reference evaluation retained the lowest specificity RMSE, 0.004369. These are conditional finite-n losses for naive accuracy ratios under the frozen FPG model; they do not establish the same ranking for latent-trait correction or a combined sensitivity–specificity objective.

In the full observed bone extraction, selective arithmetic averaging requested 37 second scans among 1,108 adults and reproduced all universal-mean labels. The five changes from the first-scan classification were the same five participants present in the ≥50-year subset, in which 36 of 517 participants were selected for a second scan; the two age scopes therefore did not supply independent replications. The evidence for preserving changed labels thus rested on five bone events, with precision estimated in the same repeat cohort. These were operational label comparisons and retrospectively emulated requirements, not verified corrections of disease status or prospective savings.

Table S4 Observed requirements for selective bone re-scanning and preservation of universal-mean label changes

Paired setting	Participants	Second measurements requested	Universal changes recovered	Missed changes
Femoral neck age ≥50 y	517	36 (7.0%)	5 / 5	0
Femoral neck age ≥20 y	1,108	37 (3.3%)	5 / 5	0

Requirements are retrospectively emulated in participants who already had paired measurements. The selective rule requests the second measurement within a fixed two-error-unit window; otherwise it retains the first label. Bone uses arithmetic-BMD averaging and an SD fixed at the threshold, with the error scale estimated in the same 1,108 pairs. Universal-mean changes are

relative to first-measurement labels, not errors against disease truth. The ≥50-year participants are nested in the ≥20-year row.

Table S5 Attainable frontier sensitivity at fixed apparent false-positive rates in 24 single-site strata

Site	Sex	Age, y	Prevalence, %	NHANES III: benchmark Se (FPR, %)	NHANES III: frontier Se at FPR 1%/5%/10%	ISCD minimum: benchmark Se (FPR, %)	ISCD minimum: frontier Se at FPR 1%/5%/10%
Femoral neck	Women	50–59	3.5	0.812 (0.48)	0.900/0.998/1.000	0.846 (0.41)	0.936/1.000/1.000
Femoral neck	Women	60–69	7.3	0.844 (0.93)	0.852/0.987/0.999	0.872 (0.79)	0.894/0.995/1.000
Femoral neck	Women	70–79	15.2	0.884 (1.71)	0.833/0.969/0.995	0.905 (1.45)	0.875/0.983/0.999
Femoral neck	Women	≥80	28.6	0.920 (2.90)	0.841/0.956/0.988	0.934 (2.43)	0.878/0.973/0.995
Femoral neck	Men	50–59	0.5	0.727 (0.09)	0.990/1.000/1.000	0.777 (0.08)	0.997/1.000/1.000
Femoral neck	Men	60–69	1.0	0.749 (0.16)	0.969/1.000/1.000	0.795 (0.14)	0.987/1.000/1.000
Femoral neck	Men	70–79	3.1	0.810 (0.42)	0.911/0.999/1.000	0.844 (0.36)	0.945/1.000/1.000
Femoral neck	Men	≥80	7.1	0.858 (0.84)	0.876/0.992/1.000	0.883 (0.71)	0.913/0.997/1.000
Total hip	Women	50–59	1.3	0.806 (0.18)	0.977/1.000/1.000	0.859 (0.14)	0.995/1.000/1.000
Total hip	Women	60–69	3.2	0.840 (0.39)	0.935/1.000/1.000	0.883 (0.31)	0.972/1.000/1.000
Total hip	Women	70–79	8.7	0.902 (0.74)	0.924/0.997/1.000	0.928 (0.57)	0.959/1.000/1.000
Total hip	Women	≥80	19.8	0.923 (1.63)	0.890/0.983/0.998	0.943 (1.25)	0.932/0.994/1.000
Total hip	Men	50–59	0.2	0.792 (0.03)	1.000/1.000/1.000	0.849 (0.02)	1.000/1.000/1.000
Total hip	Men	60–69	0.1	0.695 (0.01)	1.000/1.000/1.000	0.778 (0.01)	1.000/1.000/1.000
Total hip	Men	70–79	0.3	0.755 (0.04)	1.000/1.000/1.000	0.822 (0.03)	1.000/1.000/1.000
Total hip	Men	≥80	1.6	0.836 (0.19)	0.982/1.000/1.000	0.880 (0.15)	0.996/1.000/1.000
Lumbar spine	Women	50–59	6.1	0.893 (0.56)	0.939/0.999/1.000	0.893 (0.56)	0.939/0.999/1.000
Lumbar spine	Women	60–69	10.9	0.913 (0.90)	0.921/0.996/1.000	0.913 (0.90)	0.921/0.996/1.000
Lumbar spine	Women	70–79	17.2	0.935 (1.21)	0.924/0.994/1.000	0.935 (1.21)	0.924/0.994/1.000
Lumbar spine	Women	≥80	21.5	0.945 (1.37)	0.929/0.993/1.000	0.945 (1.37)	0.929/0.993/1.000
Lumbar spine	Men	50–59	2.3	0.870 (0.24)	0.981/1.000/1.000	0.870 (0.24)	0.981/1.000/1.000
Lumbar spine	Men	60–69	2.8	0.912 (0.21)	0.992/1.000/1.000	0.912 (0.21)	0.992/1.000/1.000
Lumbar spine	Men	70–79	2.1	0.875 (0.21)	0.986/1.000/1.000	0.875 (0.21)	0.986/1.000/1.000
Lumbar spine	Men	≥80	3.8	0.901 (0.32)	0.976/1.000/1.000	0.901 (0.32)	0.976/1.000/1.000

Se, apparent sensitivity against one DXA scan; FPR, apparent false-positive rate (1 − apparent specificity). The frontier is the apparent ROC curve of thresholding precision-error-free BMD: at each false-positive rate, no index test whose errors are independent of the DXA error given true BMD can show higher apparent sensitivity. The benchmark is the frontier point of the perfect classifier X = D (threshold T = −2.5). Consecutive sampling; single-site references; prevalence is the latent (precision-error-free) prevalence from the NHANES 2005–March 2020 distributions. Spine values coincide because both scenarios assume 1.9%. Computed post hoc (frontier_table.py); at the benchmark's own false-positive rate the calculation reproduced the saved benchmarks to within 2.2×10^{-16}.

Supplementary Methods S4. FDA-authorized devices

Sources and strata. Accuracy, DXA-defined prevalence, reference definitions and repeatability were taken from the FDA summaries of Rho (DEN230023), OsteoSight Hip (K251408) and

DeepXray Spina (K253192) and from the independent validation of Rho by Auriat et al. (main-text references 6–8 and 22). Nine FDA strata were analyzed: OsteoSight Hip (all patients, with no device result counted as negative; the subset with a device result is a sensitivity row), DeepXray Spina by sex (both sexes pooled as an additional row) and Rho by cohort (TNI6, OMN and OAI) and sex. The independent cohort contributed eight operating points: two sexes, two targets (low BMD, $T < -1.0$; osteoporosis, $T \leq -2.5$) and two Rho score cutoffs (7 and 8). The inputs, with page and table references to the source documents, are listed in device_inputs.csv and device_repeat_inputs.csv.

Prevalence matching. For each stratum, all mixture components of the sex-specific NHANES 2005–March 2020 distribution of true log BMD at the reference site were shifted by one amount so that the probability of a positive single scan equaled the reported DXA-defined prevalence. OsteoSight Hip reported prevalence only for all patients (90.7% women), so the women's distribution was used, with the pooled-sex distribution as a sensitivity analysis. The pooled DeepXray Spina row mixes the two sex-specific matched distributions with patient weights; because the device uses no demographic inputs and the probability of a positive scan given true BMD does not depend on sex, a single threshold on true BMD is the optimal comparator.

Frontier for lowest-of-sites references. For single-site references the frontier thresholds true BMD (Supplementary Methods S1). For lowest-of-sites references (Rho TNI6 and OMN: femoral neck or L1–L4; independent cohort: femoral neck, total hip or L1–L4), the most powerful rule at a fixed apparent false-positive rate thresholds the probability of a positive reference given true BMD at all sites (Neyman–Pearson lemma). This probability is one minus a bivariate or trivariate normal orthant probability of the correlated scan errors (femoral neck–total hip error correlation 0.30; spine–hip 0.25), evaluated by Gauss–Legendre integration of Plackett's identity for two sites and by conditioning on the spine error for three. Frontier points were obtained by Monte Carlo with 2,000,000 draws per stratum and a randomized test at ties, so that the achieved false-positive rate equals the device's; the Monte Carlo standard error of frontier sensitivity was at most 0.0003.

Reference share. At the device's own false-positive rate, the frontier sensitivity F is the largest apparent sensitivity that any test with errors independent of the scan's could show; $1 - F$ is the part of the shortfall attributable to the reference and F minus device sensitivity the part attributable to the device. The reference share, $(1 - F)/(1 - \text{device sensitivity})$, was computed at the reported operating point and at the four corners of the reported sensitivity and specificity confidence intervals; the smallest and largest values define the reported range.

Latent-trait correction and AUC ceiling. Where the latent-trait model describes the operating point (single-site strata with a device result), the correction of Supplementary Methods S1 was applied with effective sample sizes $n_{eff} = p(1 - p)/SE^2$, where SE is the width of the reported 95% confidence interval divided by 3.92 (Table S8). The AUC ceiling is the area under the frontier.

Repeatability. For the six repeat sets, reported device agreement and κ were compared with the model agreement of two independent DXA scans at the same threshold and DXA-positive proportion (Table S9). The reported positive proportions are those of DXA: read as the device's own positive proportions, the reported agreement would allow κ of only 0.26–0.60 in the four

Rho sets (reported 0.68–0.76) and 0.72–0.77 in the two DeepXray Spina sets (reported 0.70–0.71). The DXA flip rate is the model probability that two scans disagree; it is not an observed repeat in these cohorts. The same model gives the single-scan reference's own flip rate in the evaluation populations (Tables S6–S7).

Sensitivity analyses and checks. Sensitivity analyses used the pooled-sex distribution for OsteoSight Hip, a scale-and-location match to its reported femoral-neck BMD mean and SD, and a bilateral femoral-neck rule for Rho (Table S10). Before the single run, 14 prespecified checks reproduced the saved national benchmarks, the frontier table (Table S5) and the latent-trait round trip to within 10^{-9} and cross-checked the multisite calculations against Monte Carlo; run checks confirmed prevalence matching and the achieved false-positive rates. Inputs, script and outputs were hashed before the run.

Input caveats. DeepXray Spina and the independent cohort report image-level accuracy, and effective sample sizes come from the reported intervals; the independent cohort's intervals are not adjusted for images clustered within patients, so its ranges understate uncertainty. Independent-cohort scans had one to three sites but were modeled with three, which slightly overstates reference error. Rho's "at least one femoral neck" was modeled with one femoral neck; a bilateral variant changed benchmark sensitivity by −0.001 to −0.005 (Table S10). Rho prevalences are reported to whole percentages. The DeepXray Spina rows at $T < -1.0$ in Table S9 assume the same shift as at $T \leq -2.5$, which did not reproduce the reported $T < -1.0$ proportions, so they are illustrative.

Table S6. FDA-authorized devices against the single-scan frontier: all FDA rows and precision scenarios

Row	Prevalence	Device Se / Sp (FPR)	Benchmark Se / Sp	Frontier Se at device FPR	Reference share, ISCD (CI corners)	Share, NHANES III / best case	Device AUC / AUC ceiling	Reference false + / − per 1,000	DXA repeat κ / flip
OsteoSight Hip, all (no result counted negative)	0.667	0.370 / 0.951 (0.049)	0.970 / 0.935	0.960	6.3% (4.7%–8.2%)	8.9% / 1.3%	— / 0.994	20.1 / 21.8	0.867 / 5.9%
OsteoSight Hip, with a device result	0.667	0.441 / 0.943 (0.057)	0.970 / 0.935	0.965	6.2% (4.0%–9.4%)	8.8% / 1.2%	0.837 / 0.994	20.1 / 21.8	0.867 / 5.9%
DeepXray Spina, women	0.275	0.787 / 0.934 (0.066)	0.943 / 0.980	0.991	4.1% (1.7%–10.9%)	4.1% / 0.1%	0.957 / 0.996	15.7 / 14.3	0.894 / 4.2%
DeepXray Spina, men	0.114	0.571 / 0.942 (0.058)	0.921 / 0.991	0.999	0.3% (0.0%–6.8%)	0.3% / 0.0%	0.924 / 0.998	9.0 / 7.7	0.883 / 2.4%
DeepXray Spina, both sexes (pooled)	0.249	0.770 / 0.936 (0.064)	0.941 / 0.982	0.993	3.2% (1.3%–7.8%)	3.2% / 0.1%	0.955 / 0.997	14.6 / 13.2	0.895 / 3.9%
Rho, OAI women	0.51	0.44 / 0.90 (0.10)	0.955 / 0.953	0.984	2.8% (0.9%–6.8%)	4.6% / 0.2%	0.81 / 0.993	22.9 / 22.9	0.871 / 6.5%
Rho, OAI men	0.28	0.22 / 0.98 (0.02)	0.923 / 0.973	0.902	12.5% (6.0%–21.1%)	16.9% / 3.0%	0.79 / 0.994	21.6 / 19.3	0.857 / 5.8%
Rho, TNI6 women	0.89	0.70 / 0.90 (0.10)	0.989 / 0.902	0.989	3.6% (2.7%–4.8%)	4.8% / 0.6%	0.89 / 0.997	10.1 / 10.8	0.850 / 2.9%
Rho, TNI6 men	0.66	0.49 / 0.90 (0.10)	0.966 / 0.938	0.981	3.7% (1.9%–6.4%)	5.0% / 0.5%	0.82 / 0.993	22.6 / 21.2	0.863 / 6.2%
Rho, OMN women	0.69	0.48 / 0.94 (0.06)	0.971 / 0.938	0.970	5.8% (2.3%–14.3%)	7.4% / 1.1%	0.87 / 0.994	20.1 / 19.3	0.871 / 5.5%
Rho, OMN men	0.52	0.41 / 0.92 (0.08)	0.953 / 0.954	0.976	4.1% (1.7%–10.4%)	5.6% / 0.5%	0.83 / 0.993	24.5 / 22.0	0.869 / 6.5%

Se, sensitivity; Sp, specificity; FPR, false-positive rate (1 − Sp); CI, confidence interval. Benchmark, apparent accuracy of a perfect classifier against one scan. Frontier, the highest apparent Se at the device's FPR for any test with errors independent of the scan's. Reference share, (1 − frontier Se)/(1 − device Se), with the range over the corners of the reported CIs, at ISCD minimum precision; NHANES III and best-case shares in the next column. Reference false + / − per 1,000, people whose single scan is positive without, or negative with, true low BMD or osteoporosis in the matched population. DXA repeat κ / flip, model agreement of two independent scans in the matched population. For TNI6 and OMN (lowest of femoral neck or L1–L4) the frontier is the posterior frontier of the two-site rule (Monte Carlo SE ≤0.0002). OsteoSight Hip with a device result assumes the overall prevalence; —, device AUC not reported.

Table S7. Rho in the independent cohort: image-level operating points against the single-scan frontier

Row	Prevalence	Device Se / Sp (FPR)	Benchmark Se / Sp	Frontier Se at device FPR	Reference share, ISCD (CI corners)	Share, NHANES III / best case	Device AUC / AUC ceiling	Reference false + / − per 1,000	DXA repeat κ / flip
Women, T < −1.0, cutoff 7	0.771	0.645 / 0.841 (0.159)	0.978 / 0.931	0.994	1.7% (1.3%–2.2%)	2.6% / 0.1%	0.836 / 0.995	17.0 / 15.8	0.870 / 4.6%
Women, T < −1.0, cutoff 8	0.771	0.408 / 0.956 (0.044)	0.978 / 0.931	0.967	5.5% (4.6%–6.7%)	7.0% / 1.4%	0.836 / 0.995	17.0 / 15.8	0.870 / 4.6%
Men, T < −1.0, cutoff 7	0.667	0.383 / 0.961 (0.039)	0.966 / 0.937	0.950	8.1% (6.0%–11.5%)	10.0% / 2.2%	0.848 / 0.993	22.7 / 20.9	0.864 / 6.0%
Men, T < −1.0, cutoff 8	0.667	0.186 / 0.992 (0.008)	0.966 / 0.937	0.891	13.3% (10.6%–22.3%)	15.8% / 5.2%	0.848 / 0.993	22.7 / 20.9	0.864 / 6.0%
Women, T ≤ −2.5, cutoff 7	0.257	0.872 / 0.583 (0.417)	0.931 / 0.982	1.000	0.0% (0.0%–0.0%)	0.0% / 0.0%	0.820 / 0.996	17.8 / 13.7	0.884 / 4.4%
Women, T ≤ −2.5, cutoff 8	0.257	0.689 / 0.802 (0.198)	0.931 / 0.982	1.000	0.0% (0.0%–0.0%)	0.0% / 0.0%	0.820 / 0.996	17.8 / 13.7	0.884 / 4.4%
Men, T ≤ −2.5, cutoff 7	0.235	0.636 / 0.845 (0.155)	0.924 / 0.981	1.000	0.1% (0.0%–0.2%)	0.2% / 0.0%	0.833 / 0.995	18.0 / 14.2	0.874 / 4.5%
Men, T ≤ −2.5, cutoff 8	0.235	0.381 / 0.952 (0.048)	0.924 / 0.981	0.976	3.8% (2.4%–6.2%)	5.0% / 0.3%	0.833 / 0.995	18.0 / 14.2	0.874 / 4.5%

Abbreviations and definitions as in Table S6. Reference, lowest available femoral neck, total hip or L1–L4 T-score, modeled with three sites; Rho score cutoffs 7 and 8; posterior frontier (Monte Carlo SE ≤0.0003). Source: Auriat et al. (main-text reference 22). The source intervals are image-level and not adjusted for images clustered within patients.

Table S8. Latent-trait-corrected accuracy of the devices (ISCD minimum precision)

Row	Reported Se / Sp	Corrected Se (95% CI)	Corrected Sp (95% CI)	τ	Effective n, Se / Sp
OsteoSight Hip, with a device result	0.441 / 0.943	0.441 (0.392–0.492)	0.946 (0.923–0.962)	0.175	371 / 543
DeepXray Spina, women	0.787 / 0.934	0.796 (0.741–0.842)	0.936 (0.916–0.952)	0.069	268 / 773
DeepXray Spina, men	0.571 / 0.942	0.579 (0.372–0.761)	0.942 (0.899–0.968)	0.131	23 / 193
Rho, OAI women	0.44 / 0.90	0.442 (0.378–0.508)	0.902 (0.846–0.939)	0.220	224 / 171
Rho, OAI men	0.22 / 0.98	0.223 (0.156–0.308)	0.981 (0.958–0.991)	0.174	117 / 335

Delta-method 95% CIs with effective sample sizes derived from the reported CIs (Supplementary Methods S4). τ, SD of the index test's noise on the log-BMD scale. Rows where the latent-trait model does not describe the operating point (no device result counted negative; lowest-of-sites rules; pooled sexes) are not corrected.

Table S9. Repeatability of the devices and of one DXA scan in the device repeat sets

Repeat set	n	DXA-positive	Device agreement / κ / flip	DXA κ / flip, ISCD	DXA κ / flip, NHANES III	DXA κ / flip, best case	Flip ratio, device / DXA (ISCD)
Rho, same day, same time (bilateral images)	414	83.0%	85.0% / 0.686 / 15.0%	0.859 / 4.0%	0.841 / 4.5%	0.929 / 2.0%	3.8
Rho, same day, >1 min apart	87	83.0%	88.5% / 0.759 / 11.5%	0.859 / 4.0%	0.841 / 4.5%	0.929 / 2.0%	2.9
Rho, same day, >2 min apart	32	88.0%	84.4% / 0.690 / 15.6%	0.853 / 3.1%	0.834 / 3.5%	0.925 / 1.6%	5.0
Rho, different day within 2 years	269	89.0%	86.2% / 0.681 / 13.8%	0.850 / 2.9%	0.831 / 3.3%	0.924 / 1.5%	4.7
DeepXray Spina, same day	118	18.6%	91.5% / 0.703 / 8.5%	0.891 / 3.3%	0.891 / 3.3%	0.942 / 1.8%	2.6
DeepXray Spina, different days	76	26.3%	90.8% / 0.706 / 9.2%	0.893 / 4.1%	0.893 / 4.1%	0.944 / 2.2%	2.2
DeepXray Spina, same day, T < −1.0 (illustrative)	118	60.8% (model)	89.0% / 0.764 / 11.0%	0.893 / 5.1%	—	—	2.2
DeepXray Spina, different days, T < −1.0 (illustrative)	76	70.8% (model)	88.2% / 0.743 / 11.8%	0.889 / 4.6%	—	—	2.6

n, patients with repeat imaging; DXA-positive, the reported DXA-defined positive proportion (Supplementary Methods S4). Device agreement, κ and flip rate as reported; DXA κ and flip, model agreement of two independent DXA scans at the same threshold and DXA-positive proportion. The two DeepXray Spina rows at T < −1.0 assume the same distribution shift as at T ≤ −2.5, which did not reproduce the reported T < −1.0 proportions; they are illustrative. Bilateral-image and different-day sets include real differences between sides or over time; same-day repeats alone gave flip ratios of 2.6–5.0.

Table S10. Sensitivity analyses and derived quantities for the device analysis

(a) Bilateral femoral neck for Rho (ISCD minimum precision): benchmark with one femoral neck and L1–L4 against both femoral necks and L1–L4.

Row	One femoral neck and L1–L4: Se / Sp	Both femoral necks and L1–L4: Se / Sp	Change, Se / Sp
Rho, TNI6 women	0.9886 / 0.9020	0.9873 / 0.9140	−0.0013 / +0.0120
Rho, TNI6 men	0.9657 / 0.9375	0.9627 / 0.9456	−0.0030 / +0.0081
Rho, OMN women	0.9708 / 0.9378	0.9680 / 0.9458	−0.0028 / +0.0080
Rho, OMN men	0.9528 / 0.9541	0.9483 / 0.9601	−0.0045 / +0.0060

(b) OsteoSight Hip: distribution choice and dispersion (ISCD minimum precision). The reported femoral-neck BMD was 0.716 ± 0.146 g/cm²; the location-shift model implies 0.694 ± 0.125 g/cm²; the scale-and-location match (λ = 1.17) reproduces the reported SD.

Variant	Benchmark Se / Sp	Frontier Se at FPR 0.049	Reference share	AUC ceiling
Women's distribution, location shift (primary)	0.970 / 0.935	0.960	6.3%	0.994
Pooled-sex distribution, location shift	0.970 / 0.936	0.962	6.1%	0.994
Women's distribution, scale and location	0.974 / 0.944	0.970	4.7%	0.995

(c) Rho TNI6 to OAI, women: part of the observed sensitivity drop attributable to prevalence and the change of reference rule.

Precision scenario	Observed drop in Se	Frontier Se at FPR 0.10, TNI6 / OAI	Part attributable to the reference
ISCD minimum	0.26	0.989 / 0.984	+0.005
NHANES III	0.26	0.986 / 0.974	+0.012
Best case	0.26	0.998 / 0.999	−0.001

(d) One femoral-neck scan in screening populations (T ≤ −2.5): false labels per 1,000 screened and the share of positive scans that are truly positive.

Group	Precision scenario	True prevalence	Scan-positive prevalence	False + per 1,000	False − per 1,000	Positive scans truly positive
Women ≥65 y	NHANES III	15.8%	16.1%	17.2	14.0	0.893
Women ≥65 y	ISCD minimum	15.8%	16.0%	14.1	11.8	0.912
Women ≥65 y	Best case	15.8%	15.8%	6.5	6.0	0.959
Men ≥70 y	NHANES III	4.3%	4.5%	7.5	5.2	0.833
Men ≥70 y	ISCD minimum	4.3%	4.4%	6.0	4.4	0.863
Men ≥70 y	Best case	4.3%	4.3%	2.7	2.3	0.938

Table S11. Further practical implications for accuracy studies

Question	Evidence from this study	Recommendation
Is a reported operating point plausible?	28 of 29 audited (Se, Sp) pairs lay at or below the matched frontier, one above it at face value, and all 29 below the generous frontier; one AUC of 1.000 lay above its ceiling at face value	Check sensitivity and specificity jointly against the frontier at the study's DXA-defined prevalence; sensitivity above the fixed-status benchmark can reflect a legitimate threshold trade-off.
How large must a study be?	Estimated n for a 95% CI width ≤0.10 on latent-trait-corrected Se (extrapolated from simulated widths): consecutive women 60–69 y 4,199–5,132; consecutive men 70–79 y 10,909–12,116; 1:1 case–control 621–806	Plan precision on the number of DXA-positive participants; oversampling DXA-positives is efficient for sensitivity

Se, sensitivity; Sp, specificity; CI, confidence interval. Sample sizes are extrapolated from simulated interval widths (Supplementary Methods S1).

Table S12. Apparent accuracy of a perfect classifier against one DXA scan in US screening populations, NHANES 2005–March 2020

Population	Reference site	Latent prevalence, %	Se: NHANES III / ISCD / best	Sp	PPV	κ	Se, duplicate DXA
Adults ≥50 y	Femoral neck	5.5	0.863 / 0.888 / 0.949	0.994	0.896	0.872	0.906

Population	Reference site	Latent prevalence, %	Se: NHANES III / ISCD / best	Sp	PPV	κ	Se, duplicate DXA
Adults ≥50 y	Total hip	2.7	0.882 / 0.914 / 0.964	0.997	0.911	0.893	0.919
Adults ≥50 y	Lumbar spine	7.0	0.915 / 0.915 / 0.957	0.995	0.929	0.916	0.941
Women ≥65 y	Femoral neck	15.8	0.893 / 0.912 / 0.959	0.983	0.911	0.884	0.926
Women ≥65 y	Total hip	9.5	0.905 / 0.930 / 0.970	0.992	0.924	0.905	0.935
Women ≥65 y	Lumbar spine	16.0	0.933 / 0.933 / 0.965	0.989	0.941	0.925	0.953
Men ≥70 y	Femoral neck	4.3	0.833 / 0.863 / 0.938	0.995	0.879	0.849	0.885
Men ≥70 y	Total hip	0.7	0.812 / 0.863 / 0.943	0.999	0.871	0.839	0.872
Men ≥70 y	Lumbar spine	2.6	0.885 / 0.885 / 0.942	0.998	0.910	0.895	0.921

Se, apparent sensitivity: the share of positive scans that are truly positive; Sp, apparent specificity; PPV, the share of truly positive people whose scan is positive (the perfect classifier's apparent positive predictive value); κ, Cohen's kappa. Consecutive sampling. Each Se is the highest apparent sensitivity attainable at its own false-positive rate (1 − Sp) for these single-site references. Sp, PPV, κ and duplicate-DXA Se are at NHANES III precision. Precision CV (femoral neck/total hip/spine): NHANES III (1988–1994) 3.01/2.40/1.9%; ISCD minimum 2.5/1.8/1.9%; assumed best case 1.2/0.8/1.0%. The NHANES III spine value is assumed from ISCD. Duplicate DXA uses the mean of two log measurements. For women aged ≥65 y, the 60–69 y distribution represents ages 65–69 y, weighted by their NHANES share of the 60–69 y stratum (an approximation). Defining osteoporosis by the lowest of femoral neck, total hip and L1–L4, as in clinical diagnosis [1], raised the ≥50-year femoral-neck benchmark at NHANES III precision from 0.872 to 0.909 in women and from 0.777 to 0.856 in men.

Table S13. What the results imply for accuracy studies and device evaluations that use DXA as the reference

Question	Evidence from this study	Recommendation
How should reported accuracy be read?	A perfect classifier shows Se 0.695–0.945 (NHANES III precision) and 0.777–0.945 (ISCD minimum) across 24 strata, at false-positive rates of 0.01–2.9%. Frontier Se at a 1% false-positive rate 0.833–1.000 (Table S5); AUC ceiling ≥0.992.	Read Se and Sp together against the benchmark and frontier matched to site, population, prevalence and precision. High Se alone can reflect a lenient threshold.

Question	Evidence from this study	Recommendation
What should a device evaluation report?	Reference error explained 0.3–12.5% of the authorized devices' sensitivity shortfall at ISCD precision (0.3–16.9% at NHANES III precision). At T < −1.0 a perfect classifier shows Sp of only 0.90–0.97. Device results changed on repetition 2.2–5.0 times as often as DXA's would (2.6–5.0 in same-day repeats); a repeat DXA scan would change 2.4–6.5% of reference labels. Precision was mentioned in 13 of 136 published studies with full text.	Report the facility's DXA precision (ISCD design: 15 patients × 3 or 30 × 2 scans; CV = LSC/2.77) [1]; threshold, site rule and prevalence by sex; accuracy against the matched benchmark and frontier; how often repeating the reference and repeating the device change the result, side by side; no positive-only confirmation; and the measured CV used for any correction.
Which correction for DXA error?	Conditional-independence (Staquet) bias +0.014 to +0.112 in all 144 core scenarios, coverage 0.00–0.94. Latent-trait bias −0.003 to +0.007, coverage 0.940–0.974 at n ≥500; | bias| 0.0067–0.0085 with the CV wrong by one point; −0.032 to −0.020 when half the index response depended on binary status.	Use latent-trait correction with the scanner's measured precision. Do not use conditional-independence correction when the index test tracks bone density. Show sensitivity to the index mechanism.
How should the reference be built?	Uncorrected (four bone scenarios): confirming positives only (two-of-two) turned the Se underestimate into an overestimate; selective re-scanning halved the bias for 6.9–10.5% more scans and lowered Se RMSE in 5 of 12 budget combinations; universal duplication raised it in 11 of 12. Corrected (two CV 3% bone scenarios): repetition raised Se variance per scan (selective ×1.04–1.08; universal ×1.88–1.96).	Choose the rule with the estimator. If correcting, scan everyone once and spend repeat scans on a precision study in the study scanner. If not, avoid universal duplication and positive-only confirmation, and report the rule with its signed bias.

Se, sensitivity; Sp, specificity; RMSE, root-mean-square error; CV, coefficient of variation. Simulation values are for femoral-neck scenarios with n ≥500 unless stated. Two further rows (plausibility of a reported operating point; sample size) are given in Table S11.

Supplementary figures

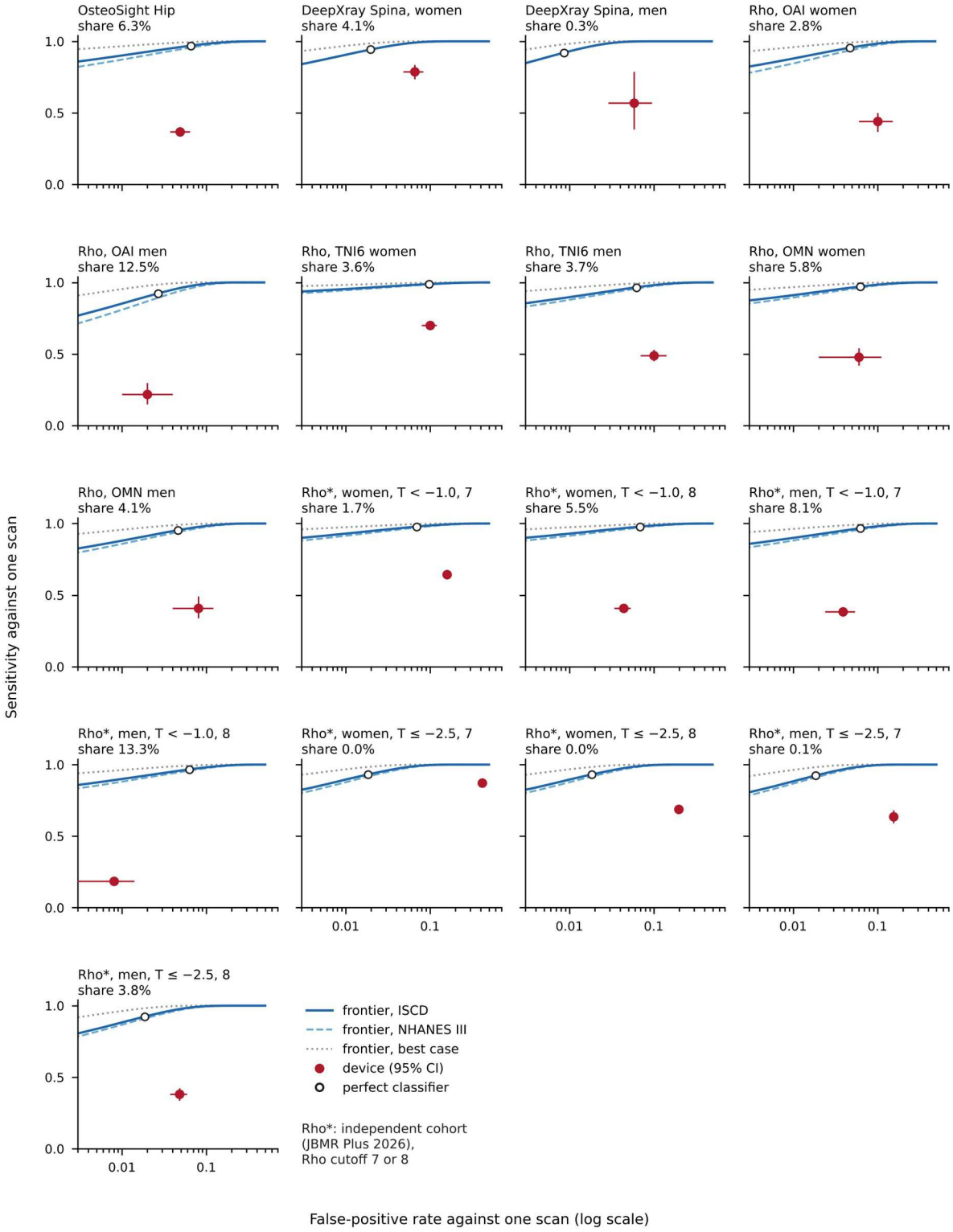


Fig. S1. Every device operating point lies below its frontier; the vertical gap is the part of the sensitivity shortfall that belongs to the device. Frontier curves at ISCD minimum (solid), NHANES III (dashed) and best-case (dotted) precision for the 17 operating points of Fig. 4c, with each device's reported sensitivity and false-positive rate (red; bars, reported 95%

confidence intervals) and the perfect classifier's point (open circle). Panel titles give the reference share at ISCD precision. Rho*, independent cohort (main-text reference 22), Rho score cutoff 7 or 8. For lowest-of-sites references the frontier is the posterior frontier (Supplementary Methods S4).

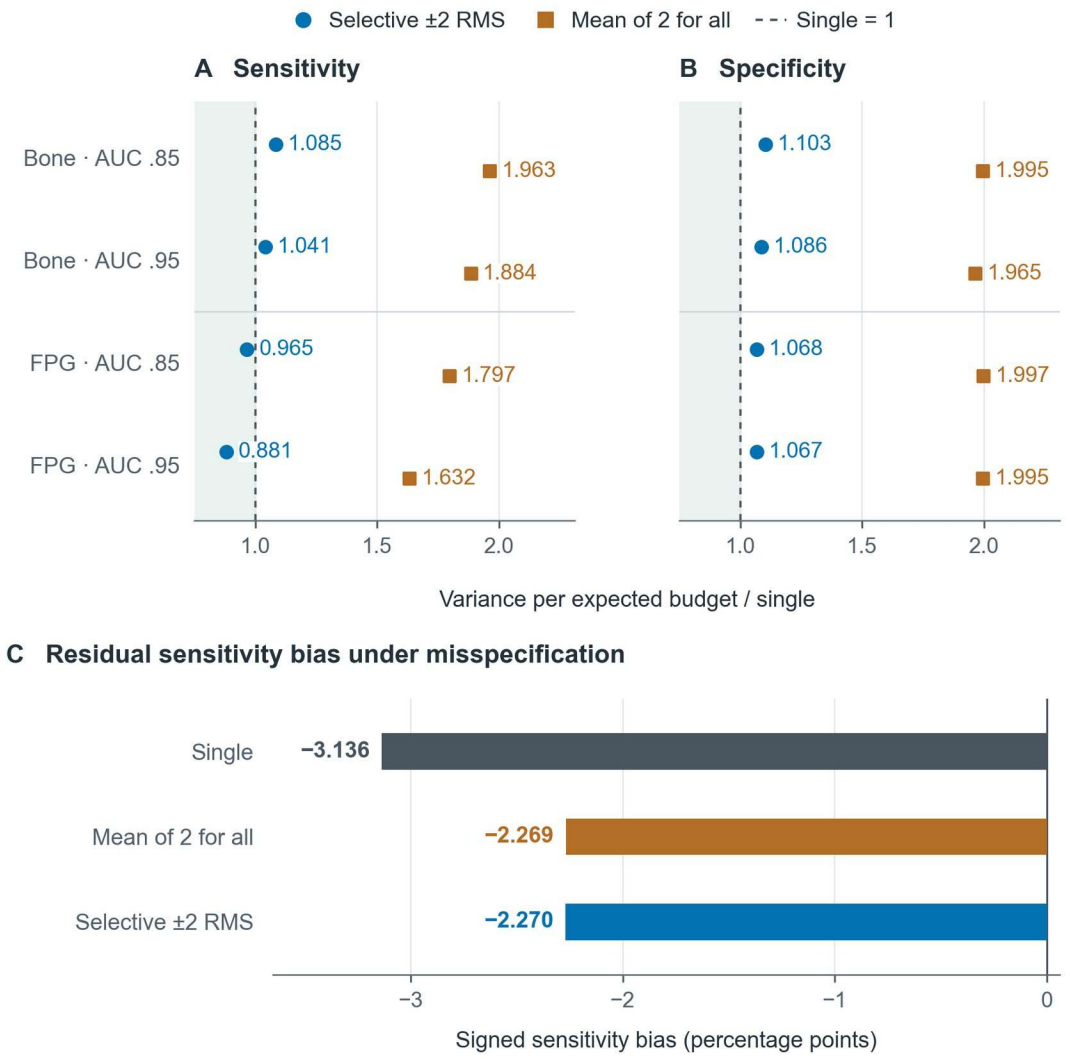


Fig. S2. After correctly specified latent-trait correction, repeating the reference raised variance per scan in both bone scenarios but lowered sensitivity variance in the glucose models. The binary latent-trait correction estimates both index parameters from apparent sensitivity and specificity, with the latent distribution and reference-error kernel fixed. (a, b) Correctly specified local asymptotic sensitivity and specificity variance per expected reference-measurement budget, relative to one measurement; a ratio below 1 favors repetition. The two bone scenarios have CV 3% and AUC 0.85 or 0.95; the two fasting plasma glucose (FPG) scenarios have the same AUC targets. The window repeats within two original error units of the threshold, using log-scale model averaging. These ratios are not finite-sample RMSE. (c) Population sensitivity bias under a mixed index mechanism in which half the index component depends on binary disease status, so the Gaussian latent-trait correction is misspecified; this panel shows residual bias only. All values are deterministic model comparisons with calibration inputs held fixed (Supplementary Methods S3).

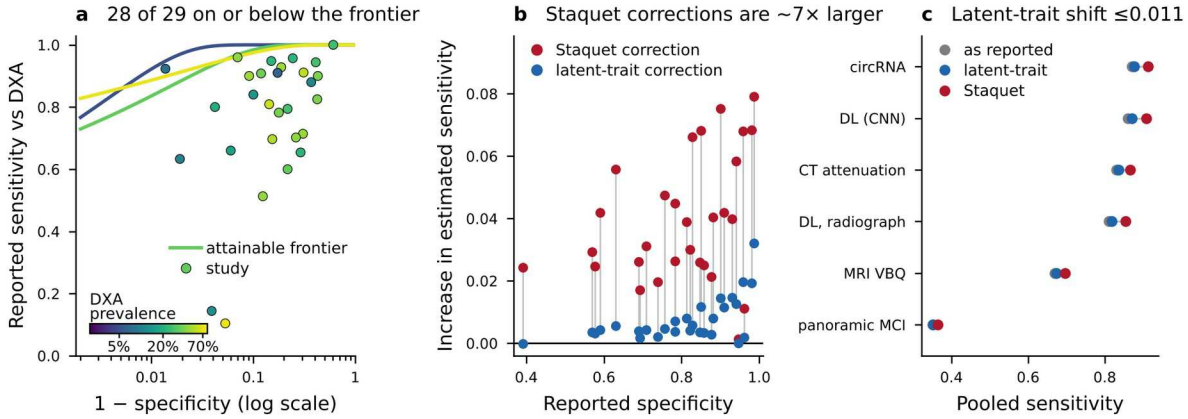


Fig. S3. Of 29 published operating points, 28 lay at or below their matched frontier and none above the most generous frontier, and correction changed published sensitivity by a median of 0.004 (latent-trait) or 0.039 (conditional independence). (a) Twenty-nine reported operating points; displayed curves illustrate pooled-sex femoral-neck frontiers at NHANES III precision for prevalences 5%, 33% and 72%. The 28/29 count uses each study's matched prevalence and the corresponding NHANES III-precision frontier; one point (sensitivity 1.000 at specificity 0.391) lies on that frontier, and the one point above it (12 of 13 at specificity 0.986) lies below the generous best-case envelope over candidate sites and sexes. No case–control study reported specificity against all participants without osteoporosis, so none is plotted. (b) Sensitivity increments from correction at ISCD precision with prevalence-matched distributions; four conditional-independence corrections took sensitivity above 1. (c) Pooled sensitivity of six meta-analyses as reported and corrected (femoral neck, prevalence 37%, ISCD precision). circRNA, circular RNA; CNN, convolutional neural network; DL, deep learning; MCI, mandibular cortical index; VBQ, vertebral bone quality.

Supplementary references

[S1] Li X, Wang J, Wang X et al (2026) CircRNAs as biomarkers for osteoporosis: a systematic review and meta-analysis. Front Endocrinol (Lausanne) 17:1768245. https://doi.org/10.3389/fendo.2026.1768245

[S2] Wulandari MD, Jorisal P, Sulay CBH, Octavius GS (2026) Diagnostic accuracy of MRI-based vertebral bone quality score for detecting low bone mass and osteoporosis: a systematic review and meta-analysis. Radiography (Lond) 32(7):103581. https://doi.org/10.1016/j.radi.2026.103581

[S3] Kinalski MA, Boscato N, Damian MF (2020) The accuracy of panoramic radiography as a screening of bone mineral density in women: a systematic review. Dentomaxillofac Radiol 49(2):20190149. https://doi.org/10.1259/dmfr.20190149