nature medicine

FOCUS | LETTERS

https://doi.org/10.1038/s41591-019-0720-z

# Automated opportunistic osteoporotic fracture risk assessment using computed tomography scans to aid in FRAX underutilization

Noa Dagan[1,2,3,6*], Eldad Elnekave[4,5,6], Noam Barda[1,2,3], Orna Bregman-Amitai[5], Amir Bar[5], Mila Orlovsky[5], Eitan Bachmat[2] and Ran D. Balicer[1,3]

**Methods for identifying patients at high risk for osteoporotic fractures, including dual-energy X-ray absorptiometry (DXA)[1,2] and risk predictors like the Fracture Risk Assessment Tool (FRAX)[3–6], are underutilized. We assessed the feasibility of automatic, opportunistic fracture risk evaluation based on routine abdomen or chest computed tomography (CT) scans. A CT-based predictor was created using three automatically generated bone imaging biomarkers (vertebral compression fractures (VCFs), simulated DXA T-scores and lumbar trabecular density) and CT metadata of age and sex. A cohort of 48,227 individuals (51.8% women) aged 50–90 with available CTs before 2012 (index date) were assessed for 5-year fracture risk using FRAX with no bone mineral density (BMD) input (FRAXnb) and the CT-based predictor. Predictions were compared to outcomes of major osteoporotic fractures and hip fractures during 2012–2017 (follow-up period). Compared with FRAXnb, the major osteoporotic fracture CT-based predictor presented better receiver operating characteristic area under curve (AUC), sensitivity and positive predictive value (PPV) (+1.9%, +2.4% and +0.7%, respectively). The AUC, sensitivity and PPV measures of the hip fracture CT-based predictor were noninferior to FRAXnb at a noninferiority margin of 1%. When FRAXnb inputs are not available, the initial evaluation of fracture risk can be done completely automatically based on a single abdomen or chest CT, which is often available for screening candidates[7,8].**

Osteoporotic fractures are a major public health concern. Over 20% of patients will not survive the year following a hip fracture[9] and an additional 20–60% will suffer residual morbidity[10]. Prophylactic interventions have been shown to decrease osteoporotic fracture risk[3,11,12], yet osteoporosis screening remains markedly underutilized. In the USA, fewer than 23% of people undergo BMD evaluation by DXA[1,2] as recommended[13]. Similarly, low utilization rates have been observed for FRAX[3–6] despite recommendations in international guidelines[13,14].

It has been suggested that this underutilization is due to lack of physician time and awareness[15–17]. FRAX could be automatically calculated when digital data are available for demographics, body measurements, diagnostic history, medication use, family medical history and life habits (smoking and alcohol use)[3]. However, most health-care providers do not have comprehensive data readily available due to the fragmented nature of medical data and high insurer turnover[18–20].

Unlike the underutilization of DXA and FRAX, CT scans are relatively ubiquitous[7,8]. CT-based metrics, such as vertebral trabecular attenuation, correlate strongly with DXA results[21,22]. Low vertebral trabecular attenuation and VCFs on CT scans have also been hailed as indicators of osteoporosis and subsequent major osteoporotic fractures[23–26].

An automated process for evaluating fracture risk based on existing or newly acquired CT scans can help in fracture risk evaluation among individuals that undergo such CT scans and opens possibilities for earlier interventions. The objective of this study was to evaluate the feasibility of creating such a CT-based fracture risk predictor and compare its performance to FRAXnb. In addition, we evaluated the performance of a FRAXnb-CT predictor that is based on inputs from both predictors.

As of 1 July 2012 (index date), there were 1,112,199 members aged 50–90 years in Clalit Health Services, a large integrated payer/provider health-care organization in Israel (see the population flowchart in Fig. 1). Of these, 30,194 (2.7%) were excluded due to lack of continuous membership. Of the remaining members, 56,197 (5.2%) had available abdomen or chest CT scans before the index date. A group of 670 (1.2%) patients were excluded because their CT was used in the bone imaging biomarker development training set; another 7,300 (13.1%) were excluded due to either the VCF or the simulated DXA T-score algorithms failing to produce a bone imaging biomarker in their available CT scans (lumbar trabecular density was deemed nonmandatory and allowed to be missing).

The success rates of the automatic algorithms in evaluating the existence of VCFs, producing the simulated DXA T-score and measuring the minimal L1-4 trabecular density were 96.5%, 84.3% and 62.3%, respectively. The overall success rate of producing the two mandatory bone imaging biomarkers (VCF and simulated DXA T-score) for a single CT was 83.6%.

The final population comprised 48,227 individuals, of whom 5,106 (10.6%) experienced a major osteoporotic fracture (hip, vertebral, proximal humerus or distal radius fractures) and 1,901 (3.9%) experienced a hip fracture before 31 June 2017, the end of the follow-up (Table 1). A total of 15.6% of the cohort had a VCF and 17.0% had a simulated T-score in the range of osteoporosis according to the algorithms. The characteristics of the CT scans of the study population that were evaluated for bone imaging biomarkers are detailed in Supplementary Table 1.

Table 2 presents the discriminatory performance of the FRAXnb, CT-based and FRAXnb-CT prediction tools for both major osteoporotic fracture and hip fracture outcomes. Table 3 presents the discriminatory performance of the CT-based and FRAXnb-CT prediction tools compared to the FRAXnb tool. (The coefficients of the CT-based tool are detailed in Supplementary Table 2.) Figure 2 presents the receiver operating characteristic curves of the three tools for the study population.

[1]Clalit Research Institute, Clalit Health Services, Tel Aviv, Israel. [2]Department of Computer Science, Ben-Gurion University, Beer Sheva, Israel. [3]School of Public Health, Ben-Gurion University, Beer Sheva, Israel. [4]Department of Diagnostic Radiology, Rabin Medical Center, Beilinson Hospital, Petah Tikva, Israel. [5]Zebra Medical Vision, Ltd, Shefayim, Israel. [6]These authors contributed equally: Noa Dagan, Eldad Elnekave. *e-mail: noada@clalit.org.il

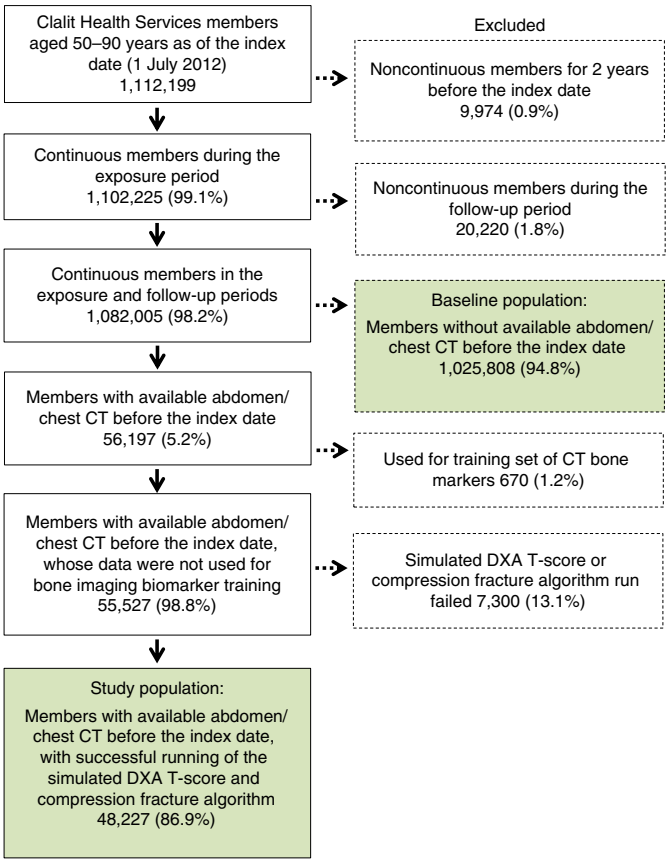


**Fig. 1 |** Flowchart of the study population, including all inclusion and exclusion criteria.

The AUC values for major osteoporotic fractures were 69.1% for FRAXnb, 70.9% for the CT-based tool and 72.3% for the FRAXnb-CT tool (Table 2). Both CT and FRAXnb-CT tools had a significantly better AUC compared to FRAXnb (+1.9% and +3.2%, respectively (Table 3)). Both tools also had improved sensitivity and PPV. A sensitivity analysis of the ability of the major osteoporotic fracture predictors to separately predict each of the four fractures that compose this outcome (Supplementary Table 3a,b) revealed that the superiority of the CT-based tool compared to the FRAXnb tool resulted from being better able to predict vertebral and proximal humerus fractures and its noninferior ability to predict hip and distal radius fractures. The superiority of the FRAXnb-CT tool compared to the FRAXnb tool resulted from a better ability to predict all fractures.

The AUC values for hip fracture prediction were 75.1% for FRAXnb, 76.0% for the CT-based tool and 77.2% for the FRAXnb-CT tool (Table 2). The AUC of the CT-based tool was noninferior to FRAXnb at a noninferiority margin of 1% and the AUC of the FRAXnb-CT tool was significantly better than FRAXnb (+2.1%) (Table 3). Sensitivity and PPV showed the same pattern of noninferiority when comparing the CT-based tool to the FRAXnb tool and superiority when comparing the FRAXnb-CT tool to the FRAXnb tool (Table 3).

The calibration performance of the three tools is presented in Fig. 3 (calibration plots) and Supplementary Table 4 (comparison between observed rates and average predicted risks). FRAXnb tended to underestimate risk, while the CT-based and FRAXnb-CT prediction tools presented better calibration. The FRAXnb-CT tool, which presented the least significant Hosmer–Lemeshow value, was used to translate the National Osteoporosis Foundation guideline[13] cutoffs into the proportion of the population at high risk.

Compared to the baseline population of 1,025,808 members without available CT scans (Supplementary Table 5), the study population had relatively older ages, greater proportion of men and higher rates of previous major osteoporotic fracture, secondary osteoporosis and glucocorticoids use. Supplementary Table 6 presents the FRAXnb discriminatory performance on this baseline population with AUCs of 71.2% and 81.1% for major osteoporotic fractures and hip fractures, respectively. A more contemporary population who underwent abdomen or chest CT scans constitutes 26.5% of 50–90-year-olds as of July 2017 and has comparable characteristics (Supplementary Table 7) to the study population, which constituted 5.2% of the 2012 population.

This study demonstrated that at least 83.6% of those aged 50–90 years who underwent routine chest or abdomen CT scans for any clinical indication could have an initial osteoporotic fracture risk evaluation performed completely automatically and solely based on the data of a single CT. We demonstrated that a CT-based risk stratification tool using three bone imaging biomarkers had comparable discriminatory performance to FRAXnb both for major osteoporotic fractures (statistical superiority) and hip fractures (statistical noninferiority).

We further showed that if the inputs of FRAXnb are available, the addition of the CT bone imaging biomarkers can produce statistically better discrimination for both outcomes. However, the clinical implications of this improvement are more pronounced for major osteoporotic fractures (+3.3% sensitivity, +0.9% PPV), and less so for hip fractures (+1.5% sensitivity, +0.1% PPV). The FRAXnb tool, which was less calibrated than the CT-based and FRAXnb-CT tools, could be recalibrated[27], but that would not change the demonstrated relative discriminatory performance.

This study used FRAXnb, the FRAX module without BMD input, for two reasons. The first was that due to DXA underutilization[2,14], mandating an existing DXA result would have resulted in a very small cohort. The second and more fundamental reason is that the intended use of this CT predictor is before a patient is assessed by DXA. Some guidelines recommend FRAXnb as a first step of risk stratification into low (reassurance), medium (further evaluation by DXA and then FRAX with BMD) and high-risk categories (treatment)[17]. However, most screening candidates are not evaluated for this first risk stratification, which is where the CT predictor could be potentially used. In countries that adhere to guidelines recommending DXA for all screening candidates (women aged 65 and older and men aged 70 and older)[13], the CT predictor could be used to identify high-risk individuals that should be pursued more proactively to make sure they undergo DXA. Thus, it should be emphasized that the CT-based predictor is not meant to replace the FRAX module with BMD, which is meant for a later phase in the risk evaluation process.

The discriminatory performance of FRAXnb in previous studies[28–30] was comparable to our baseline population and higher than that of the study population. These observations indicate that the lower performance can probably be attributed to selection bias in the more homogeneous study population[30], specifically one where all individuals underwent CT imaging. However, since CT-based fracture risk evaluation is only intended to be used on individuals undergoing CT scans, the difference between this population and the baseline population does not affect the external validity of the presented results.

Opportunistic screening for fracture risk markers using CT scans has been explored and conceptually validated. Several studies focused on simulating DXA T-scores from CT scans, but most of them compared the CT-derived metrics to DXA results and not actual fracture outcomes[22,25,31,32]. Some studies demonstrated an automated generation of CT-derived metrics and advanced the goal

**Table 1 | Characteristics of study population by FRAXnb input variables and CT-based bone imaging biomarkers**

| Input variable[a] | Mean (s.d.) | *n* (%) | Major osteoporotic fracture outcome rate, *n* (%)[b] | Hip fracture rate, *n* (%)[b] | Vertebral fracture rate, *n* (%)[b] | Proximal humerus fracture rate, *n* (%)[b] | Distal radius fracture rate, *n* (%)[b] |
|---|---|---|---|---|---|---|---|
| Overall[c] | | 48,227 (100) | 5,106 (10.6) | 1,901 (3.9) | 1,693 (3.5) | 1,035 (2.1) | 1,171 (2.4) |
| **FRAXnb input variables** | | | | | | | |
| *Age group (years)* | 69.0 (9.9) | | | | | | |
| 50–59 | | 9,937 (20.6) | 551 (5.5) | 95 (1.0) | 187 (1.9) | 124 (1.2) | 186 (1.9) |
| 60–69 | | 15,618 (32.4) | 1,209 (7.7) | 311 (2.0) | 390 (2.5) | 284 (1.8) | 343 (2.2) |
| 70–79 | | 13,947 (28.9) | 1,791 (12.8) | 662 (4.7) | 642 (4.6) | 384 (2.8) | 386 (2.8) |
| 80–89 | | 8,725 (18.1) | 1,555 (17.8) | 833 (9.5) | 474 (5.4) | 243 (2.8) | 256 (2.9) |
| *Sex* | | | | | | | |
| Women | | 25,000 (51.8) | 3,462 (13.8) | 1,196 (4.8) | 1,117 (4.5) | 775 (3.1) | 907 (3.6) |
| Men | | 23,227 (48.2) | 1,644 (7.1) | 705 (3.0) | 576 (2.5) | 260 (1.1) | 264 (1.1) |
| *Body mass index* | | | | | | | |
| Obese | | 14,559 (30.2) | 1,482 (10.2) | 440 (3.0) | 500 (3.4) | 381 (2.6) | 351 (2.4) |
| Overweight | | 18,854 (39.1) | 1,884 (10.0) | 665 (3.5) | 630 (3.3) | 391 (2.1) | 455 (2.4) |
| Normal | | 13,763 (28.5) | 1,601 (11.6) | 721 (5.2) | 526 (3.8) | 248 (1.8) | 343 (2.5) |
| Underweight | | 721 (1.5) | 109 (15.1) | 62 (8.6) | 32 (4.4) | 10 (1.4) | 14 (1.9) |
| Missing | | 330 (0.7) | 30 (9.1) | 13 (3.9) | 5 (1.5) | 5 (1.5) | 8 (2.4) |
| *Smoking* | | | | | | | |
| Nonsmoker | | 29,232 (60.6) | 3,369 (11.5) | 1,229 (4.2) | 1,099 (3.8) | 695 (2.4) | 815 (2.8) |
| Former smoker | | 11,173 (23.2) | 1,030 (9.2) | 390 (3.5) | 363 (3.2) | 215 (1.9) | 202 (1.8) |
| Current smoker | | 7,458 (15.5) | 661 (8.9) | 258 (3.5) | 218 (2.9) | 116 (1.6) | 144 (1.9) |
| Missing | | 364 (0.8) | 46 (12.6) | 24 (6.6) | 13 (3.6) | 9 (2.5) | 10 (2.7) |
| *Alcoholism* | | | | | | | |
| No | | 47,396 (98.3) | 5,010 (10.6) | 1,858 (3.9) | 1,664 (3.5) | 1,017 (2.1) | 1,148 (2.4) |
| Yes | | 831 (1.7) | 96 (11.6) | 43 (5.2) | 29 (3.5) | 18 (2.2) | 23 (2.8) |
| *Parental hip fracture* | | | | | | | |
| No | | 47,342 (98.2) | 5,046 (10.7) | 1,890 (4.0) | 1,671 (3.5) | 1,025 (2.2) | 1,150 (2.4) |
| Yes | | 885 (1.8) | 60 (6.8) | 11 (1.2) | 22 (2.5) | 10 (1.1) | 21 (2.4) |
| *Major osteoporotic fracture* | | | | | | | |
| No | | 43,183 (89.5) | 3,473 (8.0) | 1,227 (2.8) | 1,057 (2.4) | 725 (1.7) | 852 (2.0) |
| Yes | | 5,044 (10.5) | 1,633 (32.4) | 674 (13.4) | 636 (12.6) | 310 (6.1) | 319 (6.3) |
| *Secondary osteoporosis*[d] | | | | | | | |
| No | | 42,031 (87.2) | 4,298 (10.2) | 1,608 (3.8) | 1,421 (3.4) | 849 (2.0) | 983 (2.3) |
| Yes | | 6,196 (12.8) | 808 (13.0) | 293 (4.7) | 272 (4.4) | 186 (3.0) | 188 (3.0) |
| *Rheumatoid arthritis* | | | | | | | |
| No | | 46,313 (96.0) | 4,782 (10.3) | 1,779 (3.8) | 1,563 (3.4) | 967 (2.1) | 1,110 (2.4) |
| Yes | | 1,914 (4.0) | 324 (16.9) | 122 (6.4) | 130 (6.8) | 68 (3.6) | 61 (3.2) |
| *Glucocorticoids* | | | | | | | |
| No | | 42,865 (88.9) | 4,351 (10.2) | 1,645 (3.8) | 1,364 (3.2) | 879 (2.1) | 1,040 (2.4) |
| Yes | | 5,362 (11.1) | 755 (14.1) | 256 (4.8) | 329 (6.1) | 156 (2.9) | 131 (2.4) |
| **Algorithmically derived, CT-based bone imaging biomarkers** | | | | | | | |
| *Simulated T-score* | | | | | | | |
| Normal (T-score $\geq$ 1.5) | | 27,262 (56.5) | 1,831 (6.7) | 618 (2.3) | 575 (2.1) | 385 (1.4) | 442 (1.6) |
| Osteopenia ($-2.5 <$ T-score $\leq -1.5$) | | 12,769 (26.5) | 1,649 (12.9) | 625 (4.9) | 536 (4.2) | 331 (2.6) | 388 (3.0) |
| Osteoporosis (T-score $\leq -2.5$) | | 8,196 (17.0) | 1,626 (19.8) | 658 (8.0) | 582 (7.1) | 319 (3.9) | 341 (4.2) |

Continued

**Table 1 | Characteristics of study population by FRAXnb input variables and CT-based bone imaging biomarkers (continued)**

| Input variable[a] | Mean (s.d.) | *n* (%) | Major osteoporotic fracture outcome rate, *n* (%)[b] | Hip fracture rate, *n* (%)[b] | Vertebral fracture rate, *n* (%)[b] | Proximal humerus fracture rate, *n* (%)[b] | Distal radius fracture rate, *n* (%)[b] |
|---|---|---|---|---|---|---|---|
| *VCF* | | | | | | | |
| No | | 40,706 (84.4) | 3,624 (8.9) | 1,320 (3.2) | 1,015 (2.5) | 783 (1.9) | 935 (2.3) |
| Yes | | 7,521 (15.6) | 1,482 (19.7) | 581 (7.7) | 678 (9.0) | 252 (3.4) | 236 (3.1) |
| *Minimal L1-4 trabecular density*[e] | | | | | | | |
| 76th–100th percentile (137.7–536.9 HU) | | 8,335 (17.3) | 375 (4.5) | 94 (1.1) | 105 (1.3) | 95 (1.1) | 103 (1.2) |
| 51st–75th percentile (108.1–137.7 HU) | | 8,340 (17.3) | 616 (7.4) | 189 (2.3) | 171 (2.1) | 155 (1.9) | 176 (2.1) |
| 26th–50th percentile (80.5–108.1 HU) | | 8,340 (17.3) | 942 (11.3) | 338 (4.1) | 312 (3.7) | 186 (2.2) | 232 (2.8) |
| 0–25th percentile (0.1–80.5 HU) | | 8,340 (17.3) | 1,639 (19.7) | 707 (8.5) | 596 (7.1) | 305 (3.7) | 330 (4.0) |
| Missing | | 14,872 (30.8) | 1,534 (10.3) | 573 (3.9) | 509 (3.4) | 294 (2.0) | 330 (2.2) |

[a]Values within each input variable were sorted by the anticipated fracture rate, that is, the value of the variable with the lowest anticipated risk appears first. [b]Fracture rate during the follow-up period within the population of each subgroup. [c]That is, the entire study population (training + test datasets). [d]Defined by any of the following: type 1 diabetes; osteogenesis imperfecta; hyperthyroidism; hypogonadism; premature menopause; malabsorption; and chronic liver disease. [e]The 76th–100th, 51st–75th, 26th–50th and 0–25th percentiles translate to the 16–20, 11–15, 6–10 and 1–5 categories in the categorical variable, respectively. HU, Hounsfield unit.

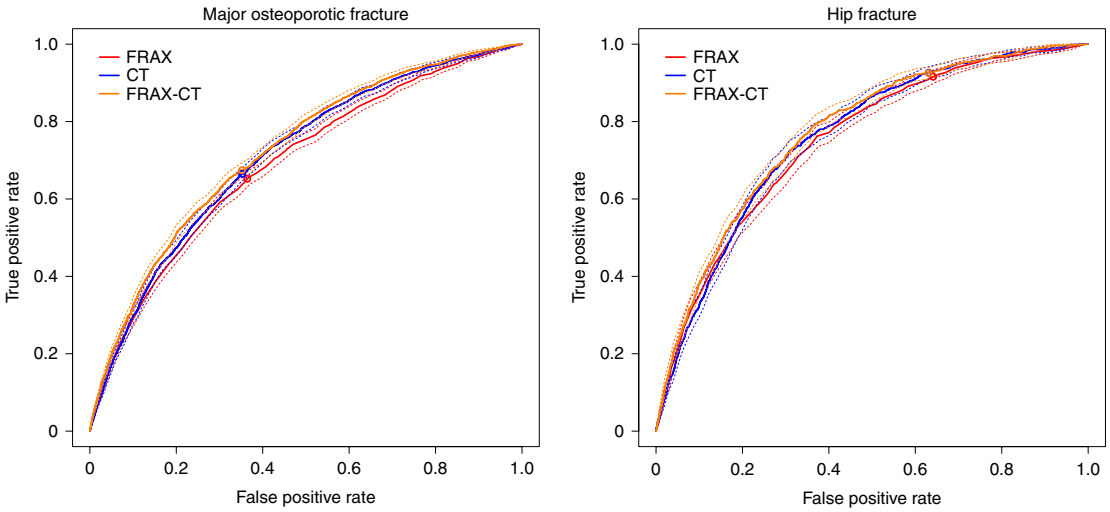


**Fig. 2 | Receiver operating characteristic curves for the major osteoporotic fracture and hip fracture outcomes.** The plot was created using the first imputed test set, which consisted of $n = 24{,}113$ individuals. The curve presents the true positive (sensitivity) and false positive rates (1 − specificity) for the different cutoffs. The circles represent the combination of true and false positive rates that corresponds to the cutoff recommended for intervention by the guidelines. The dotted lines represent the 95% CIs calculated using 500 bootstraps on the first imputed test set.

of broader population-level opportunistic screening[25,33,34]. While most studies focused on a single CT-derived feature that correlates to fracture risk, some focused on several features[23,26,35,36], including measures of bone biomechanical strength[36]. Only a minority of these studies used information regarding fracture outcomes to correlate to the CT-derived metrics[23,36], but they used case-control (204 cases, 204 controls)[23] and case-cohort (1,959 cases, 1,979 controls)[36] designs that do not preserve the true fracture incidence rate during a specified follow-up period.

The present work is aligned conceptually with these prior studies while utilizing a substantially larger dataset in a retrospective cohort design that allows for the creation of an actual fracture risk predictor. In addition, our predictor was created using multiple CT-derived metrics, all produced in a completely automatic manner. To the best of our knowledge, this is the only study of these characteristics, and the only study to compare CT-based predicted performance to a fracture predictor with accepted clinical utility. Further strengths of this study include the fact that the CT-based prediction tool is not a so-called black box. Rather, it used bone imaging biomarkers that can be interpreted and trusted by physicians and reproduced by radiologists. The ability to use both chest and abdomen CT scans was another strength because it substantially increased the scope of screening without compromising the accuracy of the prediction. Most VCFs occur in the thoracic spine and thoracolumbar junction, which are present on both scans.

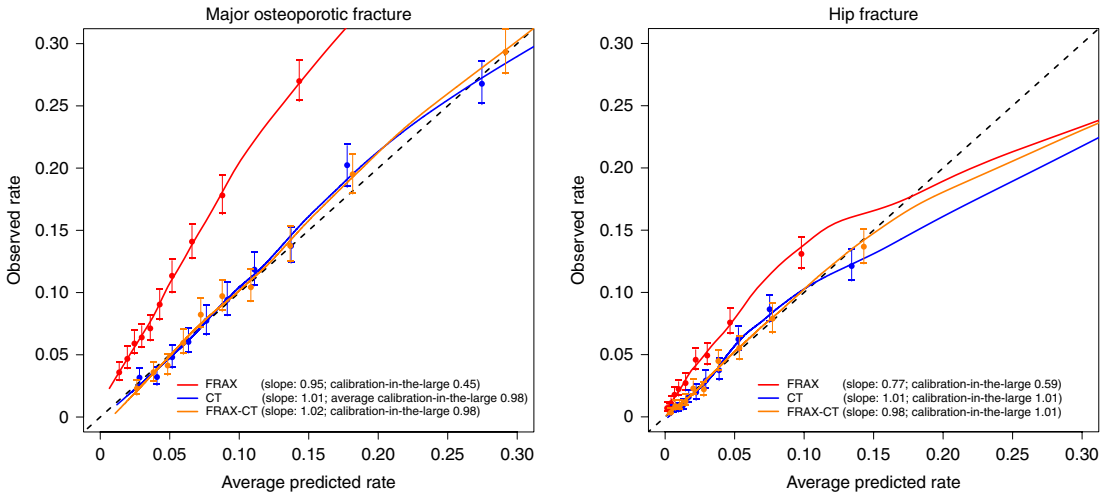


**Fig. 3 | Calibration plots for the major osteoporotic fracture and hip fracture outcomes.** The plots were created using the first imputed test dataset, which consisted of $n = 24{,}113$ individuals. Each point represents a decile of predicted risk and the corresponding observed average risk, with the points situated close to the diagonal representing good calibration. The error bars of the observed risk for each decile of the predicted risk represent the 95% CIs calculated using 500 bootstraps on the first imputed test dataset.

**Table 2 | Discriminatory performance (%) of the FRAXnb, CT-based and FRAXnb-CT prediction tools**

| Discriminatory measures[a] | FRAXnb prediction tool | CT-based prediction tool | FRAXnb-CT prediction tool |
|---|---|---|---|
| *Major osteoporotic fracture outcome* | | | |
| AUC (95% CI) | 69.1 (68.0–70.2) | 70.9 (69.9–72.0) | 72.3 (71.3–73.3) |
| Absolute risk cutoff | 4.9 | 10.3 | 10.0 |
| Sensitivity (95% CI)[b] | 64.1 (62.4–65.9) | 66.5 (64.7–68.2) | 67.4 (65.7–69.1) |
| Specificity (95% CI)[b] | 64.4 (64.2–64.7) | 64.7 (64.5–64.9) | 64.8 (64.6–65.1) |
| PPV (95% CI)[b] | 17.7 (17.0–18.5) | 18.4 (17.6–19.2) | 18.6 (17.8–19.5) |
| NPV (95% CI)[b] | 93.7 (93.4–94.1) | 94.2 (93.8–94.5) | 94.3 (93.9–94.7) |
| *Hip fracture outcome* | | | |
| AUC (95% CI) | 75.1 (73.6–76.6) | 76.0 (74.5–77.4) | 77.2 (75.7–78.6) |
| Absolute risk cutoff | 0.7 | 1.7 | 1.5 |
| Sensitivity (95% CI)[b] | 91.1 (89.3–92.9) | 92.6 (90.9–94.3) | 92.6 (90.9–94.3) |
| Specificity (95% CI)[b] | 36.8 (36.7–36.9) | 36.9 (36.8–37.0) | 36.9 (36.8–37.0) |
| PPV (95% CI)[b] | 5.6 (5.2–6.0) | 5.7 (5.3–6.1) | 5.7 (5.3–6.1) |
| NPV (95% CI)[b] | 99.0 (98.8–99.2) | 99.2 (99.0–99.4) | 99.2 (99.0–99.4) |

Analysis is based on the test dataset, which consisted of $n = 24{,}113$ individuals. The CIs were calculated using the bootstraps as detailed in the Methods. [a]All measures were evaluated and averaged across the ten imputed datasets of the test dataset. [b]For a proportion at high risk, which was set by applying the National Osteoporosis Foundation cutoffs to the FRAXnb-CT prediction tool.

**Table 3 | Comparative discriminatory performance (%) of the CT-based and FRAXnb-CT prediction tools to the FRAXnb prediction tool**

| | CT-based compared to FRAXnb | FRAXnb-CT compared to FRAXnb |
|---|---|---|
| *Major osteoporotic fracture outcome* | | |
| AUC (95% CI) | +1.9 (1.0–2.7) | +3.2 (2.6–3.8) |
| Sensitivity (95% CI) | +2.4 (0.6–4.1) | +3.3 (1.8–4.7) |
| PPV (95% CI) | +0.7 (0.2–1.1) | +0.9 (0.5–1.3) |
| *Hip fracture outcome* | | |
| AUC (95% CI) | +0.9 (−0.1 to 1.9) | +2.1 (1.5–2.7) |
| Sensitivity (95% CI) | +1.5 (−0.1 to 3.2) | +1.5 (0.2–2.8) |
| PPV (95% CI) | +0.1 (0.0–0.2) | +0.1 (0.0–0.2) |

Analysis is based on the test dataset, which consisted of $n = 24{,}113$ individuals. The CIs were calculated using bootstraps as detailed in the Methods.

This study has several limitations. First, compared to the relatively high estimated proportion of 50–90-year-olds who underwent chest or abdomen CT scans, only 5.2% of the potential study population had a relevant CT scan available. This discrepancy was due to the relatively small number of Clalit Health Services CT scanners connected to a central picture archiving and communication system (PACS) before the index date. Nevertheless, the final study population is the largest that has been studied to date to predict fracture risk based on CT scans and was large enough to provide statistical significance. Furthermore, the similarity of the study population to the broader population of screening candidates as of 2017 (26.5%) further supports the external validity of the presented results to the screening candidates' population. Another potential limitation was that the CT-based prediction tool could only be used on 83.6% of the CT scans that potentially included the relevant vertebrae (reflecting the rate of CT scans for which both VCF and simulated T-score algorithms ran successfully). The actual rate of individuals who could be evaluated for fracture risk was 86.9%, since some underwent more than one CT. This rate will potentially increase as more CT scans become available in a central PACS system[37,38].

Although potentially preventable, osteoporotic fractures are still a major cause of morbidity, mortality and health-care expenditure[39,40]. When data for automatic calculation of FRAXnb are not

available, a CT-based fracture risk predictor could be used as an initial screening tool that does not require physician time and awareness, and utilizes CT scans already performed and paid for in terms of health-care expenditure, patient time and radiation exposure. This screening method can be used on existing and newly acquired abdomen or chest CT scans, which are becoming available for a substantial percentage of screening candidates, and thus increase the number of high-risk individuals who could be identified.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of data and code availability are available at https://doi.org/10.1038/s41591-019-0720-z.

Received: 22 September 2019; Accepted: 26 November 2019;
Published online: 13 January 2020

## References

1. Curtis, J. R. et al. Longitudinal trends in use of bone mass measurement among older Americans, 1999–2005. *J. Bone Miner. Res.* **23**, 1061–1067 (2008).
2. Medical Advisory Secretariat. Utilization of DXA bone mineral densitometry in Ontario: an evidence-based analysis. *Ont. Health Technol. Assess. Ser.* **6**, 1–180 (2006).
3. Kanis, J. A., Johnell, O., Oden, A., Johansson, H. & McCloskey, E. FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporos. Int.* **19**, 385–397 (2008).
4. Marques, A. et al. The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis. *Ann. Rheum. Dis.* **74**, 1958–1967 (2015).
5. Viswanathan, M. et al. Screening to prevent osteoporotic fractures: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA* **319**, 2532–2551 (2018).
6. Beaudoin, C. et al. Performance of predictive tools to identify individuals at risk of non-traumatic fracture: a systematic review, meta-analysis, and meta-regression. *Osteoporos. Int.* **30**, 721–740 (2019).
7. Hess, E. P. et al. Trends in computed tomography utilization rates: a longitudinal practice-based study. *J. Patient Saf.* **10**, 52–58 (2014).
8. Levin, D. C., Rao, V. M. & Parker, L. Financial impact of Medicare code bundling of CT of the abdomen and pelvis. *AJR Am. J. Roentgenol.* **202**, 1069–1071 (2014).
9. Brauer, C. A., Coca-Perraillon, M., Cutler, D. M. & Rosen, A. B. Incidence and mortality of hip fractures in the United States. *JAMA* **302**, 1573–1579 (2009).
10. Dyer, S. M. et al. A critical review of the long-term disability outcomes following hip fracture. *BMC Geriatr.* **16**, 158 (2016).
11. National Institute for Health and Care Excellence. *Alendronate, Etidronate, Risedronate, Raloxifene and Strontium Ranelate for the Primary Prevention of Osteoporotic Fragility Fractures in Postmenopausal Women* (NICE, 2008); https://www.nice.org.uk/guidance/ta160/resources/raloxifene-for-the-primary-prevention-of-osteoporotic-fragility-fractures-in-postmenopausal-women-pdf-82598368491205
12. Huntjens, K. M. et al. Fracture liaison service: impact on subsequent nonvertebral fracture incidence and mortality. *J. Bone Joint Surg. Am.* **96**, e29 (2014).
13. Cosman, F. et al. Clinician's guide to prevention and treatment of osteoporosis. *Osteoporos. Int.* **25**, 2359–2381 (2014).
14. Korthoewer, D. & Chandran, M. Osteoporosis management and the utilization of FRAX®: a survey amongst health care professionals of the Asia-Pacific. *Arch. Osteoporos.* **7**, 193–200 (2012).
15. Silverman, S. L. & Calderon, A. D. The utility and limitations of FRAX: a US perspective. *Curr. Osteoporos. Rep.* **8**, 192–197 (2010).
16. Lewiecki, E. M. Managing osteoporosis: challenges and strategies. *Cleve. Clin. J. Med.* **76**, 457–466 (2009).
17. Compston, J. et al. UK clinical guideline for the prevention and treatment of osteoporosis. *Arch. Osteoporos.* **12**, 43 (2017).
18. Cebul, R. D., Rebitzer, J. B., Taylor, L. J. & Votruba, M. E. Organizational fragmentation and care quality in the US healthcare system. *J. Econ. Perspect.* **22**, 93–113 (2008).
19. Karr, A. F. et al. Comparing record linkage software programs and algorithms using real-world data. *PLoS ONE* **14**, e0221459 (2019).
20. Herring, B. Suboptimal provision of preventive healthcare due to expected enrollee turnover among private insurers. *Health Econ.* **19**, 438–448 (2010).
21. Lee, S., Chung, C. K., Oh, S. H. & Park, S. B. Correlation between bone mineral density measured by dual-energy X-ray absorptiometry and Hounsfield units measured by diagnostic CT in lumbar spine. *J. Korean Neurosurg. Soc.* **54**, 384–389 (2013).
22. Pickhardt, P. J. et al. Simultaneous screening for osteoporosis at CT colonography: bone mineral density assessment using MDCT attenuation techniques compared with the DXA reference standard. *J. Bone Miner. Res.* **26**, 2194–2203 (2011).
23. Lee, S. J., Anderson, P. A. & Pickhardt, P. J. Predicting future hip fractures on routine abdominal CT using opportunistic osteoporosis screening measures: a matched case-control study. *AJR Am. J. Roentgenol.* **209**, 395–402 (2017).
24. Melton, L. J. 3rd, Atkinson, E. J., Cooper, C., O'Fallon, W. M. & Riggs, B. L. Vertebral fractures predict subsequent fractures. *Osteoporos. Int.* **10**, 214–221 (1999).
25. Summers, R. M. et al. Feasibility of simultaneous computed tomographic colonography and fully automated bone mineral densitometry in a single examination. *J. Comput. Assist. Tomogr.* **35**, 212–216 (2011).
26. Lee, S. J. et al. Opportunistic screening for osteoporosis using the sagittal reconstruction from routine abdominal CT for combined assessment of vertebral fractures and density. *Osteoporos. Int.* **27**, 1131–1136 (2016).
27. Steyerberg, E. W. *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating* (Springer, 2009).
28. Fraser, L. A. et al. Fracture prediction and calibration of a Canadian FRAX® tool: a population-based report from CaMos. *Osteoporos. Int.* **22**, 829–837 (2011).
29. Pressman, A. R., Lo, J. C., Chandra, M. & Ettinger, B. Methods for assessing fracture risk prediction models: experience with FRAX in a large integrated health care delivery system. *J. Clin. Densitom.* **14**, 407–415 (2011).
30. Dagan, N., Cohen-Stavi, C., Leventer-Roberts, M. & Balicer, R. D. External validation and comparison of three prediction tools for risk of osteoporotic fractures using data from population based electronic health records: retrospective cohort study. *BMJ* **356**, i6755 (2017).
31. Pickhardt, P. J., Bodeen, G., Brett, A., Brown, J. K. & Binkley, N. Comparison of femoral neck BMD evaluation obtained using Lunar DXA and QCT with asynchronous calibration from CT colonography. *J. Clin. Densitom.* **18**, 5–12 (2015).
32. Ziemlewicz, T. J. et al. Opportunistic quantitative CT bone mineral density measurement at the proximal femur using routine contrast-enhanced scans: direct comparison with DXA in 355 adults. *J. Bone Miner. Res.* **31**, 1835–1840 (2016).
33. Pickhardt, P. J. et al. Population-based opportunistic osteoporosis screening: validation of a fully automated CT tool for assessing longitudinal BMD changes. *Br. J. Radiol.* **92**, 20180726 (2019).
34. Jang, S. et al. Opportunistic osteoporosis screening at routine abdominal and thoracic CT: normative L1 trabecular attenuation values in more than 20,000 adults. *Radiology* **291**, 360–367 (2019).
35. Pickhardt, P. J. et al. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Ann. Intern. Med.* **158**, 588–595 (2013).
36. Adams, A. L. et al. Osteoporosis and hip fracture risk from routine computed tomography scans: the Fracture, Osteoporosis, and CT Utilization Study (FOCUS). *J. Bone Miner. Res.* **33**, 1291–1301 (2018).
37. Sirota-Cohen, C., Rosipko, B., Forsberg, D. & Sunshine, J. L. Implementation and benefits of a vendor-neutral archive and enterprise-imaging management system in an integrated delivery network. *J. Digit. Imaging* **32**, 211–220 (2019).
38. Nagels, J., Macdonald, D. & Coz, C. Measuring the benefits of a regional imaging environment. *J. Digit. Imaging* **30**, 609–614 (2017).
39. Burge, R. et al. Incidence and economic burden of osteoporosis-related fractures in the United States, 2005–2025. *J. Bone Miner. Res.* **22**, 465–475 (2007).
40. Hernlund, E. et al. Osteoporosis in the European Union: medical management, epidemiology and economic burden. A report prepared in collaboration with the International Osteoporosis Foundation (IOF) and the European Federation of Pharmaceutical Industry Associations (EFPIA). *Arch. Osteoporos.* **8**, 136 (2013).

## Methods

**Study design.** In this retrospective cohort study, we created a CT-based prediction tool that calculated fracture risk scores automatically based on data taken solely from chest or abdomen CT scans. The tool was created by integrating three bone imaging biomarkers generated by deep learning algorithms and adding CT metadata of patient age and sex. The bone imaging biomarkers included the presence of VCFs, CT-derived simulated DXA T-scores and evaluated lumbar trabecular density. We then compared whether the performance of the CT-based tool was comparable to that of FRAXnb. In addition, we developed and explored the performance of a combined FRAXnb-CT prediction tool, which used inputs from the previous two tools, to explore whether the CT bone imaging biomarkers could further improve the performance of FRAXnb.

These three prediction tools (CT-based, FRAXnb and FRAXnb-CT) were compared for the two outcomes assessed by the FRAXnb tool: hip fractures and major osteoporotic fractures. The latter is a composite of hip, vertebral, proximal humerus or distal radius fractures.

The tools' performance was evaluated by comparing the calculated fracture risk as of 1 July 2012 (index date) to fractures occurring until the end of June 2017 (5-year follow-up period).

**Setting.** This study was performed using data from Clalit Health Services, an Israeli health-care organization that insures and provides primary, specialty and inpatient health-care services to nearly 4.5 million members, over half of the Israeli population. Israeli residents are eligible to choose a health fund as part of the country's universal health-care coverage, but switching between funds is relatively uncommon (less than 2% annually), allowing for longitudinal follow-up with low numbers lost to follow-up[41].

The study was conducted in collaboration between the Clalit Health Services research institute, a non-for-profit organization, and Zebra Medical Vision, a private sector entity. Zebra Medical Vision trained the deep learning algorithms to create three bone imaging biomarkers. Zebra Medical Vision were provided ID-encrypted CT images while remaining blinded to other clinical or outcome data about the study cohort, including information regarding osteoporotic fractures during the follow-up period. To evaluate the contribution of these markers to osteoporotic fracture prediction, the Clalit Health Services research institute crossed these markers with demographic and clinical data to create fracture prediction tools and compare their performance.

**Study population.** The study population consisted of Clalit Health Services members aged 50–90 years as of the index date, as in the FRAXnb derivation cohort[3]. To ensure the availability of sufficient clinical information, the study population had to have two years of continuous Clalit Health Services membership before the index date and through the follow-up period or until death. The study population was further required to have previously undergone an adequate abdomen or chest CT scan as of the index date. Each participant was only included once, even if multiple CT scans were available. CT scans that were included in the training of the bone imaging biomarkers or were technically inadequate for interpretation by the algorithms were excluded. Due to the retrospective nature of the study, all qualified members were included and no sample size calculation was done. The study received approval from the Institutional Review Board Committee of Clalit Health Services for studies of outpatients (study ID 0176-15-COM2). Institutional Review Board Committee approval included an exemption regarding obtaining written informed consent.

A comparison baseline population that comprised all study population candidates who had available abdomen or chest CT before the index date was used to describe the derivation cohort in terms of characteristics and FRAXnb performance. In addition, a population of Clalit Health Services members aged 50–90 years with a history of abdomen or chest CT scans as of July 2017 was used to determine if the study population was representative of future screening candidates.

**Data sources.** Demographic and clinical data from Clalit Health Services electronic health records and CT scans from Clalit Health Service PACS were used for this study. The electronic health records include sociodemographic information, diagnoses from community and hospital settings, registries, laboratory results, medication use and clinical markers (for example, body mass index (BMI) and smoking status).

**Variable definition.** Definitions for the extraction of the FRAXnb input variables and fracture outcomes, including relevant codes and categorizations, have been described in detail by Dagan et al.[30].

The algorithmic vertebral analysis was generated automatically by deep learning from abdomen and chest CT scans previously acquired for any clinical indication, which included at least the L1 vertebral body. The first bone imaging biomarker included a binary indication of VCF in any thoracolumbar vertebra (whichever of T3-L4 that were included in the scan)[42]. The second bone imaging biomarker was a numeric variable of a simulated DXA T-score of the visualized L1–L4 lumbar vertebrae[43,44]. The third bone imaging biomarker represented an evaluation of the minimal trabecular density measured in the L1–L4 vertebrae, which was categorized into 20 equally large bins, along with a missing category for CT scans where the algorithm failed to produce a result.

Multiple imputation was conducted using the functions outlined by van Buuren et al.[45] to complete any missing documentation of BMI or smoking status before the index date (the only variables for which missing data could be identified). The imputation process was repeated ten times, creating ten full datasets.

**Generation of the VCF bone imaging biomarker.** To develop the VCF detection algorithm, Zebra Medical Vision created an initial dataset of 3,701 CT examinations of the chest and/or abdomen of individuals over the age of 50 years. All CT scans were reviewed by two expert radiologists, who assessed them for the presence of VCFs as defined by the criteria outlined by Genant et al.[46]. A third radiologist (E.E.) served to arbitrate in cases where consensus could not be reached. Of the 3,701 CT studies, 2,681 (72%) were negative for the presence of VCF and 1,020 (28%) were VCF-positive ($VCF^+$); all VCFs were annotated. The presence of discogenic and end plate degenerative changes was noted in the tagging process, but these findings were not annotated visually. Similarly, the presence of traumatic or oncogenic fractures was noted and documented in the tagging process but not annotated for training.

This initial data curation process yielded substantial differences in the composition of $VCF^+$ versus VCF-negative ($VCF^-$) subgroups. The $VCF^+$ CT subgroup consisted of 61% women with an average age of 73 years (s.d. 12.4) and 39% men with an average age of 66.8 years (s.d. 16.8). The $VCF^-$ subgroup consisted of 47% women with an average age of 56.7 years (s.d. 17.4) and 53% men with an average age of 56.1 years (s.d. 17.9). Although these observations were consistent with the epidemiology of osteoporosis, the differences in demographic characteristics could introduce biases with unintended results. Training a classifier on this dataset might, for example, result in an algorithm that distinguishes between the spines of old women and younger men. Subsequently, a more demographically balanced subset of CT studies was derived including 1,673 CT studies, of which 849 were $VCF^-$ and 824 $VCF^+$. The $VCF^-$ and $VCF^+$ groups were age- and sex-matched. The data were split into training (85%) and test (15%) groups. All algorithmic training and hyperparameter tuning was performed on the training set; the held-out portion was reserved for testing.

From the axial series of an abdominal or chest CT scan, the vertebral column center line was localized on a secondarily reconstructed coronal maximum intensity projection. A single two-dimensional image was generated by tracking the midline in the volume along the sagittal plane, thereby correcting for any right/left scoliosis and rendering a virtual sagittal image. This was performed to obtain consistent sagittal views of CT examinations, most of which did not contain primary multiplanar reconstructions. Connected component analysis, thresholding and morphological operations were used to detect the posterior border of the vertebral column on the sagittal view and extract fixed size ($32 \times 32$) craniocaudal patches.

To classify a sequence of patches corresponding to a single CT series, first a patch-based binary classification convolutional neural network (CNN) was trained using the architecture described in greater detail elsewhere[42]. Then, the series of patches was fed into the patch-based classifier, resulting in a series of probabilities assigned by the model. The probabilities were then fed into a long short-term memory layer with 128 cells, followed by a single fully connected layer to aggregate the results into a single CT-series level result, trained via the cross-entropy loss.

**Generation of the simulated DXA T-score and lumbar trabecular density bone imaging biomarkers.** As described in detail elsewhere[43], the DXA simulation training and validation by Zebra Medical Vision utilized 1,843 pairs of CT and DXA scans obtained from the same individual at a 6-month interval. All imaging was performed as part of clinical practice between 2010 and 2014 within Clalit Health Services. The study population consisted of 70.8% women and 29.2% men between 50 and 80 years of age. An additional 610 CT abdomen/pelvis studies were used to achieve multiclass L1–L4 vertebral segmentation.

Spinal segmentation was derived from primary axial reconstructions, in the manner described earlier. To achieve lumbar multiclass segmentation, a cascade of two U-Nets[47] was used. Binary segmentation of the vertebral column was performed using the aforementioned virtual sagittal reconstruction. Multiclass segmentation was then performed on the basis of coregistration of each vertebral body from the virtual sagittal and coronal maximum intensity projection reconstructions, allowing reference of the vertebral body in relation to the ribs, with the first lumbar vertebral body designated as the one caudal to the last rib-bearing vertebra. This allowed for consistent identification of L1 through L4 even in the presence of lumbar variations, including four or six non-rib-bearing vertebrae. Unlike patch-based segmentation, fully CNNs add upsampling layers to standard CNNs, allowing better recovery of spatial resolution. To compensate for the resolution loss induced by pooling layers, fully CNNs introduce skip connections between their downsampling and upsampling paths.

L1–L4 trabecular attenuation values were obtained from a two-dimensional area one-third of the distance between vertebral cortices at the sagittal midline. To identify optimal pixel range intensity for T-score simulation, linear regression was used as detailed previously[43]. X-ray anteroposterior acquisition was simulated, excluding pixels outside the learned intensity range, to provide a summation map.

Then, the DXA information was used to evaluate the learned T-score regression results for each L1–L4 vertebra. Contrast and noncontrast examinations were treated in the same manner since it was previously demonstrated that administration of intravenous contrast has less effect on the attenuation of the trabecular bone in the lumbar spine[26,48], especially in a population of individuals over the age of 40 years[34].

**Calculation of fracture prediction scores.** As described in detail previously[30], we used the FRAXnb ten-year probability charts calibrated for Israel (https://www.sheffield.ac.uk/FRAX/), which were converted to five-year probabilities. The justification for this transformation has been described previously[30]. The FRAX module used was the one that does not include BMD input.

To train the CT-based and FRAXnb-CT-based predictors, the study population was randomly divided into training and test datasets (50:50 ratio). Both models were developed on the training dataset using logistic regression. The process was repeated separately in the ten imputed training datasets; the coefficients were then averaged to create a final model. A 95% confidence interval (CI) for the coefficients was estimated by applying Rubin's rule[27,49].

The CT-based prediction tool was developed using the input of the three bone imaging biomarkers, along with the age and sex variables. The FRAXnb-CT prediction tool was developed using the same inputs as the CT-based prediction tool, in addition to a single variable that represents the FRAXnb linear prediction component (the $\beta^T X$ component in which $\beta$ represents the coefficients and X represents the input variables; this component is calculated by transforming the FRAXnb predicted probability into logits[27]).

**Evaluating fracture model performance.** The evaluation of all performance measures was done only on the test dataset, which was imputed separately. The overall discriminative ability of each prediction tool was evaluated using the AUC. Additional discriminatory measures, including sensitivity, specificity, PPV and NPV were also evaluated. Since these measures are cutoff-specific, whereas the models are not all calibrated equally, the comparison between models was based on the same percentile of the study population that would be considered high risk. To choose an appropriate percentile, the best-calibrated model was used to evaluate the percentage of the population that would be considered high risk based on the National Osteoporosis Foundation guidelines, that is, 20 and 3% risk for major osteoporotic fractures and hip fractures in 10 years, respectively[13] (translated to 10 and 1.5% risk in a 5-year follow-up period).

In addition, to further characterize the discriminatory performance of the major osteoporotic fracture prediction tools (FRAXnb, CT and FRAXnb-CT), we performed a sensitivity analysis of their ability to separately predict each of the four outcomes that compose the major osteoporotic fracture outcome—hip, vertebral, proximal humerus or distal radius fractures.

Each tool's calibration was evaluated by comparing the average predicted risk with the observed percentage of major osteoporotic fractures and hip fractures during the follow-up period, stratified by deciles of predicted risk. Additionally, several measures that evaluated the overall calibration were calculated: the Hosmer–Lemeshow goodness-of-fit test; calibration slope; and calibration-in-the-large[27,50].

All performance measures were calculated separately on each of the imputed datasets and averaged to create the final performance measures. A 95% CI for tool-specific performance measures, as well as for the differences between tools, was calculated using Rubin's rules for variance estimation in multiple imputed datasets[27,49]. This was done by taking into account both the variance of 500 bootstrap samples randomly drawn with replacement within each imputed dataset and the variance of the 10 averaged performance measures between the imputed datasets. The 95% CI for the evaluation of the difference between tools was used to compare the CT-based and FRAXnb-CT tools to the FRAXnb tool, which represented the baseline performance. A 95% CI that did not include zero was used to represent a significant difference (that is, superiority), while a 95% CI indicating that the measure was no more than 1% less than the baseline performance (that is, noninferiority margin[51]) was considered to represent noninferiority.

Receiver operating characteristic curves and smoothed calibration plots were created using the first imputed dataset. Analyses were conducted using R v.3.5.2 (MICE v.3.3.0[45], ROCR v.1.0-7[52] and rms v.5.1-2[50] packages).

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The study protocol can be shared upon request. Access to the data used for this study can be made available upon request, subject to an internal review by N.D. and R.D.B. to ensure that participant privacy is protected, and subject to completion of a data sharing agreement, approval from the institutional review board of Clalit Health Services and institutional guidelines and in accordance with the current data sharing guidelines of Clalit Health Services and Israeli law. Pending the aforementioned approvals, data sharing will be made in a secure setting, on a per-case-specific manner from the chief information security officer of Clalit Health Services. Please submit such requests to N.D. (noada@clalit.org.il).

## Code availability

Requests for the statistical code will be considered by the authors according to the stated need and dependent on specific approval by the information security office of Clalit Health Services.

## References


41. Gross, R., Rosen, B. & Chinitz, D. Evaluating the Israeli health care reform: strategy, challenges and lessons. *Health Policy* **45**, 99–117 (1998).
42. Bar, A., Wolf, L., Bergman Amitai, O., Toledano, E. & Elnekave, E. Compression fractures detection on CT. In *Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis* (eds Armato, S.G. 3rd & Petrick, N. A.) 1013440 (SPIE, 2017).
43. Krishnaraj, A. et al. Simulating dual-energy X-ray absorptiometry in CT using deep-learning segmentation cascade. *J. Am. Coll. Radiol.* **16**, 1473–1479 (2019).
44. Bregman-Armitai, O. & Elnekave, E. Systems and methods for emulating DEXA scores based on CT images. Patent no. WO2016013005A2 (2019); https://patentimages.storage.googleapis.com/aa/dd/d7/a9ac0a3b551f72/WO2016013005A2.pdf
45. van Buuren, S. & Groothuis-Oudshoorn, K. MICE: Multivariate imputation by chained equations. R package version 2.22 https://cloud.r-project.org/web/packages/mice/index.html (2014).
46. Genant, H. K., Wu, C. Y., van Kuijk, C. & Nevitt, M. C. Vertebral fracture assessment using a semiquantitative technique. *J. Bone Miner. Res.* **8**, 1137–1148 (1993).
47. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-assisted Intervention (Springer, Cham.) 234–241 (2015).
48. Pickhardt, P. J. et al. Effect of IV contrast on lumbar trabecular attenuation at routine abdominal CT: correlation with DXA and implications for opportunistic osteoporosis screening. *Osteoporos. Int.* **27**, 147–152 (2016).
49. Marshall, A., Altman, D. G., Holder, R. L. & Royston, P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med. Res. Methodol.* **9**, 57 (2009).
50. Harrell, F. E. Jr. rms: Regression modeling strategies. R package version 4.4-0 https://rdrr.io/cran/rms/ (2015).
51. Walker, E. & Nowacki, A. S. Understanding equivalence and noninferiority testing. *J. Gen. Intern. Med.* **26**, 192–196 (2011).
52. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: Visualizing the performance of scoring classifiers. R package version 1.0-7 https://rdrr.io/cran/ROCR/ (2015).


## Acknowledgements

We thank S. Krispin for her assistance in editing and reviewing the manuscript. This study was supported by a grant given to Clalit Health Services and Zebra Medical Vision by the Israel Innovation Authority (Ministry of Social Equality), to promote digitalized transformation in health care (grant number 64727). The Israel Innovation Authority did not have any active involvement in the study process.

## Author contributions

N.D., N.B., E.E., E.B. and R.D.B. conceived and designed the study. N.D. extracted the data and performed all the statistical analysis. N.D. and N.B. interpreted the data. N.D. and E.E. drafted the manuscript. O.B.A., A.B. and M.O. contributed the formation of the image processing algorithms. All authors critically revised the manuscript. R.D.B. and E.B. supervised the study and are the guarantors.

## Competing interests

E.B. has no interests to disclose. N.D., N.B. and R.D.B. report having received grants from the Israel Innovation Authority during the conduct of this study. The parent company of Clalit Research Institute (N.D., N.B. and R.D.B.) owns a minority share in Zebra Medical Vision. E.E., O.B.A., A.B. and M.O. report personal fees from Zebra Medical Vision during the conduct of this study. In addition, E.E. and O.B.A. have a patent for emulating DXA scores based on CT images (WO2016013005A3) and a patent to predict osteoporotic fracture risk (WO2016013004A1).

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41591-019-0720-z.

**Correspondence and requests for materials** should be addressed to N.D.

**Peer review information** Jennifer Sargent was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

nature research

Corresponding author(s): Noa Dagna

Last updated by author(s): Nov 22, 2019

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | SQL server |
| Data analysis | Analyses were conducted using R 3.5.2 (mice, ROCR, and rms packages). |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The study protocol could be shared upon request. Requests for sharing of statistical code will be considered by the authors according to stated reason for the request and the security measures that will be suggested to protect the code. Raw patient data cannot be shared for reasons of patient privacy.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

# Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | This is a retrospective cohort analysis, based on existing medical records. Methods are strictly quantitative. |
| Research sample | The population of the study is the entire patient population of a large health fund operating in Israel matching the inclusion criteria of the baseline model used in the paper (FRAX). Accordingly, patients aged 50-90 years (as of the index date of July 1, 2012.) of both sexes were included, who had an available abdomen or chest CT scan as of the index date. |
| Sampling strategy | Due to the retrospective nature of this study all qualified members were included and no sample-size calculation was done. The final study population is the largest that has been studied to date to predict fracture risk based on CT scans, and was large enough to provide statistical significance. |
| Data collection | All data used in this study is based on existing electronic medical records and CT scans prior to an index date. No dedicated data collection was performed. |
| Timing | All data available as of an index of July 1, 2012 was used to create prediction scores. Data regarding the outcome was collected during a follow-up period of 5 years (until June 2017). |
| Data exclusions | As detailed in figure 1, study participants were included due to the following reasons: lack of continuous membership in the health fund, no available CT prior to the index date, CT used in the train set of the original algorithm development. |
| Non-participation | Participants that did not have continuous membership during the follow-up period were also excluded (a total of 20,220 which consist of 1.8% of the study candidates). |
| Randomization | This is an observational study, no randomization was performed. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |
| ☒ | ☐ Clinical data |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | The study population consisted of Clalit members aged 50-90 years (females and males) as of the index-date (July 1, 2012), with two years of continuous Clalit membership before the index-date and through the 5-year follow-up period or until death. In addition, all members were required to have an available abdomen or chest CT scan prior to the index-date. |
| Recruitment | Participants were not recruited but rather selected from a retrospective Clalit medical database. Due to the requirement to have an available abdomen or chest CT prior to the index date this is a selected population that does not represent the entire 50-90 year olds population. However, since CT-based fracture risk evaluation is only intended to be used on individuals undergoing CT scans, the difference between this population and the general baseline population does not affect the external validity of the presented results. |
| Ethics oversight | This study was approved by the Clalit research ethics committee for studies of outpatients (study ID: 0176-15-COM2). |

Note that full information on the approval of the study protocol must also be provided in the manuscript.