Machine Learning to Predict Osteoporotic Fracture Risk from Genotypes.

Vincenzo Forgetta,[1*] Julyan Keller-Baruch[2*], Marie Forest[1], Audrey Durand[3], Sahir Bhatnagar,[1] John Kemp[4,5], John A Morris[1], John A Kanis[6], Douglas P. Kiel[7], Eugene V McCloskey[8], Fernando Rivadeneira,[9] Helena Johannson[6], Nicholas Harvey[10,11], Cyrus Cooper,[10,11,12] David M Evans[4,5], Joelle Pineau[3], William D Leslie[13], Celia MT Greenwood[1,2,14,15], J Brent Richards[1,2,16]

[1]Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada,
[2]Department of Human Genetics, McGill University, Montréal, Québec, Canada,
[3]School of Computer Science, McGill University, Montréal, Québec, Canada,
[4]University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland, Australia,
[5]MRC Integrative Epidemiology Unit, University of Bristol, UK,
[6]Centre for Metabolic Bone Diseases, University of Sheffield, Sheffield, UK, and Australian Catholic University, Melbourne, Australia
[7]Institute for Aging Research, Hebrew SeniorLife, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Broad Institute of MIT & Harvard University, Boston, USA
[8]Mellanby Centre for Bone Research, Centre for Integrated Research in Musculoskeletal Ageing, University of Sheffield, and Sheffield Teaching Hospitals Foundation Trust, Sheffield, UK,
[9]Department of Internal Medicine, Erasmus Medical Center, Rotterdam, Netherlands
[10]Medical Research Council Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK
[11]National Institute for Health Research Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton, UK,
[12]National Institute for Health Research Oxford Biomedical Research Centre, University of Oxford, Oxford, UK,
[13]Department of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada,
[14]Gerald Bronfman Department of Oncology, McGill University, Montréal, Québec, Canada,
[15]Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montréal, Québec, Canada,
[16]Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom.

*These authors contributed equally to this study.

Corresponding Author:
Brent Richards, MD, MSc
Professor of Medicine McGill University
Senior Lecturer, King's College London (Honorary)

Contact:
Pavillon H-413, Jewish General Hospital
3755 Cote Ste Catherine
Montreal, QC, Canada, H3T 1E2
brent.richards@mcgill.ca

Background: Genomics-based prediction could be useful since genome-wide genotyping costs less than many clinical tests. We tested whether machine learning methods could provide a clinically-relevant genomic prediction of quantitative ultrasound speed of sound (SOS)—a risk factor for osteoporotic fracture.

Methods: We used 341,449 individuals from UK Biobank with SOS measures to develop genomically-predicted SOS (gSOS) using machine learning algorithms. We selected the optimal algorithm in 5,335 independent individuals and then validated it and its ability to predict incident fracture in an independent test dataset (N = 80,027). Finally, we explored whether genomic pre-screening could complement a UK-based osteoporosis screening strategy, based on the validated tool FRAX.

Results: gSOS explained 4.8-fold more variance in SOS than FRAX clinical risk factors (CRF) alone (r^2 = 23% vs. 4.8%). A standard deviation decrease in gSOS, adjusting for the CRF-FRAX score was associated with a higher increased odds of incident major osteoporotic fracture (1,491 cases / 78,536 controls, OR = 1.91 [1.70-2.14], P = 10^{-28}) than that for measured SOS (OR = 1.60 [1.50-1.69], P = 10^{-52}) and femoral neck bone mineral density (147 cases / 4,594 controls, OR = 1.53 [1.27-1.83], P = 10^{-6}). Individuals in the bottom decile of the gSOS distribution had a 3.25-fold increased risk of major osteoporotic fracture (P = 10^{-18}) compared to the top decile. A gSOS-based FRAX score, identified individuals at high risk for incident major osteoporotic fractures better than the CRF-FRAX score (P = 10^{-14}). Introducing a genomic pre-screening step into osteoporosis screening in 4,741 individuals reduced the number of required clinical visits from 2,455 to 1,273 and the number of BMD tests from 1,013 to 473, while only reducing the sensitivity to identify individuals eligible for therapy from 99% to 95%.

Interpretation: The use of genotypes in a machine learning algorithm resulted in a clinically-relevant prediction of SOS and fracture, with potential to impact healthcare resource utilization.

Research in Context

Evidence Before this Study

Genome-wide association studies have identified many loci associated with risk of clinically-relevant fracture risk factors, such as SOS. Yet, it is unclear if such information can be leveraged to identify those at risk for disease outcomes, such as osteoporotic fractures. Most previous attempts to predict disease risk from genotypes have used polygenic risk scores, which may not be optimal for genomic-prediction. Despite these obstacles, genomic-prediction could enable screening programs to be more efficient since most people screened in a population are not determined to have a level of risk that would prompt a change in clinical care. Genomic pre-screening could help identify individuals whose risk of disease is low enough that they are unlikely to benefit from screening.

Added Value of this Study

Using a large dataset of 426,811 individuals we trained and tested a machine learning algorithm to genomically-predict SOS. This metric, gSOS, had performance characteristics for predicting fracture risk that were similar to measured SOS and femoral neck BMD. Implementing a gSOS-based pre-screening step into the UK-based osteoporosis treatment guidelines reduced the number of individuals who would require screening clinical visits and skeletal testing by approximately 50%, while having little impact on the sensitivity to identify individuals at high risk for osteoporotic fracture.

Implications of all of the Available Evidence

Clinically-relevant genomic prediction of heritable traits is feasible using the machine learning algorithm presented here in large sample sizes. Genome-wide genotyping is now less expensive than many clinical tests, needs to be performed once over a lifetime and could risk stratify for multiple heritable traits and diseases years prior to disease onset, providing an opportunity for prevention. The implementation of such algorithms could improve screening efficiency, yet their cost-effectiveness will need to be ascertained in subsequent analyses.

Introduction

Genomic prediction of complex traits and diseases could improve the efficiency of health care systems if it can accurately identify those at risk for disease. This potential arises because genotypes are stable over the lifetime so that high risk individuals might be identified early in life, providing an opportunity for prevention years before disease onset. Further, genotypes need to be measured only once over a lifetime and could provide risk prediction for many heritable diseases. Finally, since the cost of genome-wide genotype is now reasonably affordable (approximately $40 in a research context) and falling faster than Moore's law,[1] genotype-based prediction has become less expensive than some clinical tests. For these reasons genome-wide genotyping has already been undertaken in several health care contexts.[2,3] Yet despite this promise, it remains unclear if genomics can render clinically-relevant predictions and perhaps more importantly, how such information could be used to improve care.

The use of genetic prediction can be explored using osteoporosis as an example, since it is among the most common diseases and is diagnosed primarily through bone density measures, which are highly heritable (50-85%) and highly polygenic.[4–7] Osteoporosis is defined as a reduction in bone mass and microarchitectural integrity resulting in an increased risk of fracture.[8] Many guideline-based screening strategies[9–12] for prevention of osteoporosis-related fractures incorporate the Fracture Risk Assessment Tool (FRAX),[13,14] a validated method to risk stratify individuals for treatment. Further, a FRAX-based screening program has also been shown to reduce hip fracture incidence.[15] However, population-based screening programs identify a small proportion of the screened population that is eligible for lifestyle or treatment interventions and require relatively expensive BMD testing in a large proportion of the population who will not ultimately receive treatment. Thus, current osteoporosis care models provide the opportunity to explore how genomic prediction may provide clinical utility.

At present, there is little evidence that genotype-based prediction algorithms provide clinical utility for common disease. Since the variance explained by genotypes in complex traits and diseases by genotypes has been typically low, the resulting performance of genetic prediction has not rendered important improvements over clinical risk factors.[16,17] One way to improve the variance explained by genotypes is through the use of larger sample sizes, applying genome-wide association study (GWAS) polygenic risk scores,[18–21] since for many complex traits and diseases the effect sizes of any one associated genetic risk factor is small, but taken together, the variance explained can be substantial.[22] Machine-learning algorithms may also be applied to larger sample sizes to improve prediction, where these methods learn from training data and then assess performance of the algorithm in independent test datasets.[23]

In this study, we trained a machine learning algorithm and separate polygenic risk scores on a large cohort of 341,449 individuals of British ancestry from UK Biobank with genome-wide genotypes to predict bone quantitative ultrasound-derived speed of sound (SOS) a measure that predicts osteoporotic fracture partly independently of DXA-derived bone mineral density (BMD) and clinical risk factors.[24] We then selected the optimal prediction model in 5,335 independent individuals and finally tested the clinically-relevant performance characteristics of genomically-predicted SOS (gSOS) for fracture prediction in 80,027 independent individuals. Finally, we explored how gSOS could be incorporated into FRAX-based screening strategies in order to identify a proportion of the population not needing screening and/or BMD testing.

Methods

Overall Study Design
This study included three phases (**Figure 1**). The first was to use machine learning methods[25] and polygenic risk scores to predict SOS in the training dataset, selecting the top-performing model in the model selection dataset, which we refer to as gSOS. The second phase used the independent test dataset to assess the ability of gSOS to predict fracture and the third was to explore how gSOS may be used to improve screening programs.

Cohort
UK Biobank is a large-scale health resource that follows 502,628 volunteer participants in the United Kingdom.[26] Participants within the UK Biobank have been genome-wide genotyped using Affymetrix arrays,[27] followed by genotype imputation to the Haplotype Reference Consortium.[28] UK Biobank has ethical approval from the Northwest Multi-centre Research Ethics Committee, and informed consent was obtained from all participants.

*SOS and BMD Measurement (Details in **Supplement**)*
Details of SOS measurement are available from UK Biobank (goo.gl/A5DTdP). All analyses used SOS standardized to a mean of zero and standard deviation of one. We chose to model SOS, since it was available for more individuals than BMD, has a gradient of risk for incident major osteoporotic fractures similar to BMD and a recent meta-analysis described its effect on fracture risk, allowing for the development of a SOS-based FRAX model.[24] Dual energy x-ray absorptiometry BMD was measured at the femoral neck (and is referred to hereafter as "BMD") in a subset of the population (N = 4,741), that was reserved for the test dataset in order to assess osteoporosis screening strategies.

Development of machine learning model to predict SOS (**Supplement** and **Figure 1**)
I. Training, Model Selection and Test datasets: To develop and test our prediction models, we followed best practices in machine learning to ensure unbiased estimates of model performance. Participants in UK Biobank with White British ancestry, measured SOS, and genotyping information (N= 426,811) were randomly assigned to either a training dataset (80% of participants, restricted to those without BMD), a model selection dataset (1.25%, also without BMD), or the test dataset (18.75% of participants including all with BMD) (**Figure 1** & **Table 1**).
II. GWAS: Using the training dataset (N=341,449 individuals with White British ancestry), we tested the additive allelic effects of each of the 13M SNPs passing QC, separately, on SOS using a series of linear mixed-model[29], adjusting for age, sex, assessment centre, genotyping array, and the first 20 principal components derived from the analysis of White British ancestry. Linkage disequilibrium-independent associations where obtained using PLINK by clumping SNPs in linkage equilibrium at a $r^2 > 0.05$ and selecting a single most significant SNP from within each clumped set.
III. Machine learning models: We fit 6 regularized linear machine learning models[25] based on L1-penalized least absolute shrinkage and selection operator (LASSO) regression[25] to the training dataset to predict SOS using only SNPs with P-values smaller than a chosen set of thresholds (**Table S2**). The model selection dataset was then used to optimize model performance by choosing the best regularization parameter (λ) which controls the number of SNPs contributing to the prediction. Finally, the estimated model for the SNP-set giving the smallest root mean square error in the model selection dataset was taken forward for testing in the test dataset. Hence, we ensured that the performance of only one final model from the machine learning strategy was evaluated in the test dataset. We refer to this final predictor as "gSOS", in which SOS is predicted only from genotypes. All evaluations of the role of gSOS on screening tests were performed in the test dataset.

IV. Polygenic Risk Scores: Traditional polygenic risk scores[18] were generated summarizing different sets of SNPs, selected by P-value threshold, as described in the **Supplement**.

*Measurement of FRAX risk factors (Details in **Supplement**).*
The FRAX risk score for fracture can be generated with or without BMD, referred to BMD-FRAX and clinical risk factor CRF-FRAX, respectively. FRAX CRFs were assessed at the baseline visit and included age, sex, body mass index (BMI), previous fracture, smoking, corticosteroid use, rheumatoid arthritis and secondary causes of osteoporosis. More than three daily units of alcohol and parental history of hip fracture variables were not available and were set to "no" for the entire population. Age was recorded at baseline visit, and sex was self-reported and verified by genotype. Individuals with discordant sex between self-report and genotype were excluded.

*Fracture Definitions (Details in **Supplement**)*
Site-specific major osteoporotic fractures[30] were defined using questionnaire data from UK Biobank and ICD10 codes for hospital admissions. For ICD10 codes (**Table S1**), only primary diagnoses were considered, since the date of secondary diagnoses was not consistently available. The date of ICD-10 identified fractures was determined from administrative data and if dated after the baseline visit was labelled as incident. Prevalent fractures were identified as those reported by questionnaire or from ICD10 codes that occurred prior to the baseline visit. Trauma type was not included since mechanism of trauma is variably captured by ICD codes and even high trauma fractures are predictive of future low trauma fractures and are associated with low BMD.[31,32]

Generation of FRAX scores
Four FRAX scores were generated for each person in the test dataset: 1) CRF-FRAX; 2) CRF plus measured-SOS FRAX (termed SOS-FRAX); 3) CRF plus gSOS FRAX (termed gSOS-FRAX) and 4) CRF plus BMD FRAX (termed BMD-FRAX). CRF-FRAX and BMD-FRAX were generated using the FRAX algorithm.[33] To generate the SOS-FRAX probabilities, we took the CRF-FRAX score (as pre-test odds) and updated this using measured SOS or gSOS (gradient of risk per standard deviation decrease of 1.42 for major osteoporotic fractures and 1.80 for hip fracture from the largest meta-analysis to date which did not include UK Biobank).[24] We used this method since a SOS-FRAX score is currently not available.

Non-White British, Population Stratification and Cryptic Relatedness
The training, model selection and test datasets included only White British participants to reduce population stratification. The GWAS was also controlled for the top 20 principle components of ancestry to reduce effects of cryptic relatedness. To test the generalizability of gSOS and assess for potential effects of population stratification and cryptic relatedness we tested gSOS on fracture outcomes in a smaller cohort of non-White British participants (defined in the **Supplement**).

Testing of Prediction Performance in the Test Dataset
Prediction performance was first tested by assessing the correlation between gSOS and measured SOS in the test dataset. Next, the correlations between gSOS and FRAX CRFs were assessed. gSOS was tested for association with risk of incident major osteoporotic fracture and hip fracture, comparing this to measured SOS and BMD measured at the femoral neck using univariate and multivariable logistic regression, including CRF-FRAX as a covariate. The association of each of the four FRAX scores with risk of incident major osteoporotic fracture was also tested in the test dataset by first log-transforming the FRAX scores to correct for a skewed distribution and then using logistic regression. The area under the receiver operator curve was

calculated for each of the FRAX scores, and these were compared to the CRF-FRAX. The same procedure was then completed for incident hip fracture.

Since health care systems may want to use genomics-based predictors to identify an extreme of the population at risk of disease, we next divided the distribution of gSOS into percentiles and contrasted the bottom decile with the top decile to describe the associated increased risk of fracture. Since 10% of any health care system is a large number of people, we then explored odds of fracture for those in the lowest 25th quantile (lowest 4% of the distribution), compared to the top 25th quantile.

Genomic Pre-Screening

Health care providers may also want to use genomics-based predictors to better stratify individuals for inclusion in screening programs, assuming that genome-wide genotyping has been already undertaken. We therefore explored if gSOS could be used to identify individuals who would be unlikely to benefit from screening for osteoporosis, because their risk of fracture would be low. To do so, we implemented the UK National Osteoporosis Guideline Group (NOGG) approach within the test dataset.[10] The NOGG Guidelines aim to identify individuals at risk for fracture in a cost-efficient manner by reserving clinical visits and BMD testing for those at increased risk.

We first evaluated the effectiveness of NOGG Guidelines to appropriately risk stratify individuals for therapy, comparing this to a gold standard where all individuals had BMD testing and were risk stratified by BMD-FRAX. We chose NOGG over other national guidelines since a NOGG-like approach was recently demonstrated in a randomized controlled trial to be potentially effective in reducing hip fractures.[15] Further, the NOGG Guidelines recommend less BMD testing than those of the US National Osteoporosis Foundation[11] and Osteoporosis Canada.[34] Consequently an improvement in BMD testing efficiency over NOGG would suggest improvements in other guideline settings. NOGG-guidelines utilise 10-year absolute risks of fracture using FRAX and suggest treatment or reassurance based on thresholds of risk, which are age dependent and consider competing risks. These 10-year absolute risks were calculated for all individuals for major osteoporotic fracture and participants were risk stratified as per NOGG Guidelines (**Figure S4**), which aims to reduce the number of BMD tests by restricting such tests to those at moderate risk. We then tested the sensitivity of NOGG guidance to identify individuals at high risk, comparing NOGG guidance to a BMD-FRAX gold-standard. Since BMD is required to calculate BMD-FRAX, the genomic pre-screening was evaluated only in individuals from the test dataset with BMD measures (N = 4,741).

In order to implement genomic pre-screening, we designated all individuals over a defined threshold of gSOS to be reassured, and not to undergo a clinical visit for CRF-based FRAX testing, or BMD testing (**Figure S5**). We chose the threshold of gSOS which maximally reduced clinical visits and BMD testing, while minimizing the loss of sensitivity to identify individuals who would be recommended for treatment by the gold standard. The performance of pre-screening was then assessed by counting the number of individuals who would require a clinical visit and BMD testing and calculating its sensitivity to correctly assign participants to reassurance or treatment.

Results

Cohort Characteristics

Table 1 describes the FRAX risk factors for the training, model selection and test data sets. There was no clinically-relevant difference in any of the osteoporosis-related risk factors, as expected, since these sets were generated randomly. As planned, all individuals with BMD measures were included in the test dataset. There were little differences in demographics or CRFs among individuals with BMD measured, except for a higher prevalence of smoking and incident major osteoporotic fractures.

GWAS

After quality control, 13,958,249 SNPs were included in the GWAS. The GWAS in the training set identified 1,404 independent ($r^2 \leq 0.05$) genome-wide significant loci at a P-value threshold of $< 5 \times 10^{-8}$ **Figure S1** shows the Manhattan and QQ plots for this GWAS.

Variance Explained in SOS in the Model Selection Set

Age, sex and BMI explained 4.0% of the variance in SOS. Adding the remaining FRAX clinical risk factors increased the variance explained to 4.8%. Polygenic risk scores across different P-value thresholds explained at most 18.5% (95% confidence interval [CI]: 16.6-20.4%) of the variance in SOS (**Table S2** and **Figure 2**). The machine-learning algorithm improved the variance explained in SOS to a maximum of 25.0% (23.0-27.0%) (**Table S2** and **Figure 2**). All six SNP-sets using the machine learning algorithm outperformed polygenic risk scores substantially. **Figure S2** provides detailed information on the optimal algorithm tuning parameters.

Variance Explained in SOS in the Test Set

The top model from the model selection set was then tested for its correlation with SOS in the test set. This model (machine learning algorithm set at P-value $\leq 10^{-4}$, including 21,717 activated SNPs, which are those SNPs retained by the machine learning algorithm) explained 23.2% (95% CI: 22.7-23.7%) of the variance in measured SOS (**Table S2** and **Figure 2**). We then designated this model as "gSOS", which was subsequently assessed for performance characteristics in the test dataset. **Figure S3** demonstrates that, as expected, the estimated effects of the activated SNPs from the machine learning algorithm were attenuated, when compared to the effects estimated from the GWAS

gSOS Correlation with other Clinical Risk Factors and BMD

Since gSOS was derived only from genotypes it should be independent of factors that are not in the causal pathway between SNPs and SOS. We found that while measured SOS was correlated with all measured FRAX clinical risk factors, gSOS was not (**Table 2**). These findings suggest that if gSOS is associated with risk of fracture it would be also independent of these other clinical risk factors.

Prediction of Incident Major Osteoporotic and Hip Fracture

There were 1,491 incident major osteoporotic fracture cases and 78,536 incident fracture-free controls in the test dataset (**Figure 3**). We also identified 209 hip fracture cases and 79,818 incident fracture-free controls in the same dataset. In univariate models, decreased SOS, gSOS and BMD were all strongly associated with increased odds of incident fracture, however, among these predictors, gSOS had the highest risk per standard deviation decrease. Since the predictive value of these skeletal measures may be reduced after accounting for correlated FRAX CRFs, we tested the association of SOS, gSOS and BMD with incident major osteoporotic and hip fractures, adjusting for CRF-FRAX. We found that association of fracture

with SOS and BMD were more attenuated by inclusion CRF-FRAX (**Figure 3**), likely because gSOS is not associated with FRAX CRFs, while SOS and BMD are. Associations with incident hip fractures showed similar effect sizes, but with wider confidence intervals. There were too few incident hip fractures amongst the 4,741 participants with BMD measures to report meaningful findings.

Incorporating gSOS into FRAX Probability of Fracture

We next tested the performance of the four FRAX probabilities on risk of incident major osteoporotic fracture. We found that a standard deviation increase in each FRAX score on the log scale was strongly associated with incident major osteoporotic fracture (**Table 3**). The magnitude of association was lower when using CRFs alone and increased when introducing any of the three skeletal measures. The area under the receiver operator curves improved significantly, with the incorporation of skeletal measures when compared to CRFs alone, as has been shown previously.[33] **Table 3** shows that the area under the receiver operator curve for gSOS-FRAX was equivalent to that for BMD-FRAX but was lower than SOS-FRAX.

We next generated the FRAX probabilities using hip fracture as an outcome. **Table 3** shows that each of the models was strongly associated with odds of incident hip fracture and both SOS-FRAX and gSOS-FRAX improved the area under the curve when compared to CRF-FRAX. The area under the curve for SOS-FRAX and gSOS-FRAX was equivalent.

Extremes of gSOS

Comparing individuals in the bottom decile of the gSOS distribution to those in the top decile we found that those in the bottom were at a 3.25-fold increased risk of major osteoporotic fracture (95% CI: 2.50-4.25, P = 10^{-18}). Since a decile represents a large absolute number of individuals from any health care system we next tested the risk associated with being in the bottom 25th of the gSOS distribution, compared to those in the top 25th of the distribution and found a ~5-fold increased risk of incident major osteoporotic fracture (OR = 4.94 [3.07-7.91], P = 10^{-11}).

Non-White British

To assess for generalizability of gSOS and possible effects of population stratification and cryptic relatedness we tested its performance in non-White British participants. There were 494 incident major osteoporotic fracture cases and 42,902 controls amongst non-White British participants. There were 69 incident hip fracture cases and 43,327 controls. The results for non-White British participants were very similar to those amongst White British participants (**Table S3**). gSOS was associated with a higher odds of major osteoporotic and hip fracture, both with and without FRAX CRFs. The area under the receiver operator curve was improved using gSOS-FRAX compared to CRF-FRAX for major osteoporotic and hip fracture (P = $1x10^{-4}$ and P = $4x10^{-4}$, respectively). Quantiles of gSOS also demonstrated a high odds of fracture (bottom 10th quantile OR = 3.71 [95% CI: 2.33-5.89, P = $1x10^{-8}$]; bottom 25th quantile OR = 8.17 [95% CI: 3.22-20.75, P = $5x10^{-6}$]).

Genomic Pre-Screening

Table 4 demonstrates the implementation of current NOGG Guidelines for screening amongst all 4,741 UK Biobank participants with BMD testing. **Figure S4** shows what would be the clinical outcomes (i.e. reassurance or treatment recommendation) based on current NOGG Guidelines. The sensitivity for correct treatment allocation, compared to undertaking BMD-FRAX on all individuals as the gold-standard, was 99.1% and the specificity was 99.6%. To achieve this performance, 2,455 of the 4,741 individuals required a clinical visit while 1,013 required a BMD test (**Figure S4** and **Table 4 and Table S4**).

We then tested whether pre-screening with gSOS could improve efficiency of these guidelines, by reassuring everyone over a certain gSOS threshold. Reassuring all NOGG-eligible participants at a high gSOS threshold would result in little improvement in efficiency. Alternatively reassuring all participants above too low a gSOS threshold would result in a large number of participants being reassured who would actually be recommended for treatment if they had had their BMD-FRAX measured. Thus, we selected the threshold of gSOS that would minimize the number of BMD tests not changing care, but also not increase the number of participants who are eligible for therapy but would be reassured (**Figure S5**). **Figure S6** and **Table 4** demonstrate that gSOS pre-screening would reduce the number of participants requiring a clinical visit from 2,455 to 1,273 and decrease the number of BMD tests from 1,013 to 473. Using gSOS-based pre-screening, the sensitivity for correct treatment allocation decreased from 99.1% to 94.7% (**Table 4**), while the specificity, positive predictive value and negative predictive value were 99.9%, 98.6% and 99.5%, resptively.

Discussion

Here we have developed and validated a machine learning algorithm to predict a skeletal measure of fracture risk (gSOS) from genotypes alone. gSOS has performance characteristics for fracture risk that are similar to measured SOS and femoral neck BMD. The machine learning algorithms used here provide improved prediction over traditionally-used polygenic risk scores. While the exact use of such prediction algorithms in clinical care will require further investigation, we have explored several potential scenarios

First, in multivariate models, gSOS was more strongly associated with major osteoporotic fracture than was SOS or BMD. For fracture prediction, gSOS outperformed FRAX CRFs alone. The lower tail of the gSOS distribution was at a substantially increased risk of incident fracture, providing the opportunity to identify individuals at elevated risk. gSOS pre-screening was able to identify a large proportion of the population for whom osteoporosis screening programs would not result in changes in clinical care. These findings suggest that genomic prediction can render clinically-relevant information that could be used to improve health care delivery systems. We suggest that such advances will not be limited to osteoporosis, but rather, given large enough sample sizes and use of machine learning algorithms, genomic prediction will become clinically-relevant for several heritable traits and diseases.

Already at least seven large health care delivery systems have invested in genome-wide genotyping of a large proportion of their population, upon whom electronic health record (EHR) data are available.[2,3] Since the costs associated with genome-wide genotyping have now dropped below those of many routine clinical tests, the use of genomic prediction of clinically relevant traits and diseases will likely encourage other health care systems to make similar investments—provided genomic predictors can be shown to be clinically relevant and cost efficient.

The fracture prediction performance of gSOS is higher than expected given the amount of variance explained in measured SOS. This is likely for at least three reasons. First, gSOS is not subject to measurement error inherent in SOS; second it is independent of other FRAX clinical risk factors and last, genotypes are measured with very high precision.[35] It is not currently clear which machine learning methods are best suited to prediction. Nonetheless, the observed association between genotypes and continuous traits has largely been linear, making multivariate regression-based methods, like LASSO, attractive.

Screening programs for osteoporosis are expensive, with estimates of approximately $50,000 USD per quality adjusted life year gained,[36] but are less expensive using NOGG guidance.[37] Current guidelines suggest screening for a large proportion of the population,[9–11] yet most patients are not identified to have a clinically-actionable level of risk.[15,38] This provides an opportunity for genetically-derived measures of risk to increase cost-efficiencies, particularly in health care systems where investments have been made in genome-wide genotyping. While the health-economic realities of such prediction are unclear at present, we have provided one example of how gSOS could be used to improve the efficiency of screening based on the NOGG Guidelines. We note that the efficiencies of gSOS pre-screening are likely magnified in the context of UK Biobank, where most subjects are generally healthy, yet those receiving BMD tests were more likely to be smokers and have had an incident fracture.

Previous attempts to predict osteoporosis from genomic data did not substantially increase discrimination, when compared to standard clinical measures alone, likely because the GWAS that underpinned these attempts were derived from 32,961 individuals and explained only 5.8% of the variance in BMD.[17,39] The improvement in variance explained in in this study was likely due to the increase in sample size afforded by UK Biobank as well as the ability of the machine learning algorithm to estimate the joint contributions of all SNPs rather than their independent effects. Other attempts to predict BMD have been based on several dozen genome-wide significant SNPs,[17] whereas here we have used a machine learning algorithm, which has enabled the consideration of approximately up to 642,127 SNPs. Recently, a machine-learning algorithm was used to predict estimated BMD, but from a GWAS sample size that was one third of that used here.[40] An additional prediction algorithm using incorrect definitions of osteoporosis was able to explain only 17% of the variance in estimated BMD and was able to predict fracture only marginally better than chance alone.[41] Therefore the data presented here represent a clinically-relevant improvement in genomic prediction.

The generalizability of genomic prediction can be influenced by population stratification and cryptic-relatedness. Encouragingly, gSOS, despite being developed in the White British subset of UK Biobank, performed similarly in the non-White British subset, who have a different ancestry than the White British subset.

This study has important limitations. While gSOS does risk stratify for fracture, it must be emphasized that this is a probability and is not deterministic—such probabilities have similar characteristics to other risk factors for common disease. BMD would have been a more natural skeletal measure of fracture risk, since this is more often used in the clinic. We anticipate that the same methods here will be useful for BMD genomic prediction, once appropriately large sample sizes are available and emphasize that the predictive performance of gSOS is similar to that for measured BMD. While nearly all FRAX CRFs were available for study, parental hip fracture history was not. It is possible that the clinical relevance of gSOS may be attenuated after accounting for this heritable risk factor, and this will need to be empirically tested once such data are available. gSOS-FRAX could, in the future be optimized to include competing risks and interaction effects, as has been done for BMD-FRAX. Our results have not been validated in an external sample but have been internally validated in 80,027 individuals in UK Biobank. Like participants in most cohort studies, UK Biobank participants are, on average, healthier than the general population.[42] Thus, external validation in a truly population-based study may provide helpful estimates of predictive abilities of gSOS.

In summary, we have developed and validated a machine-learning algorithm to predict SOS from genotypes alone, which when predicting fracture, performs similarly to measured SOS and BMD. The methods and results provided here may help to guide genomic prediction programs

for other clinically-relevant risk factors and diseases, providing new opportunities for improved clinical care through genomics-enabled medicine.

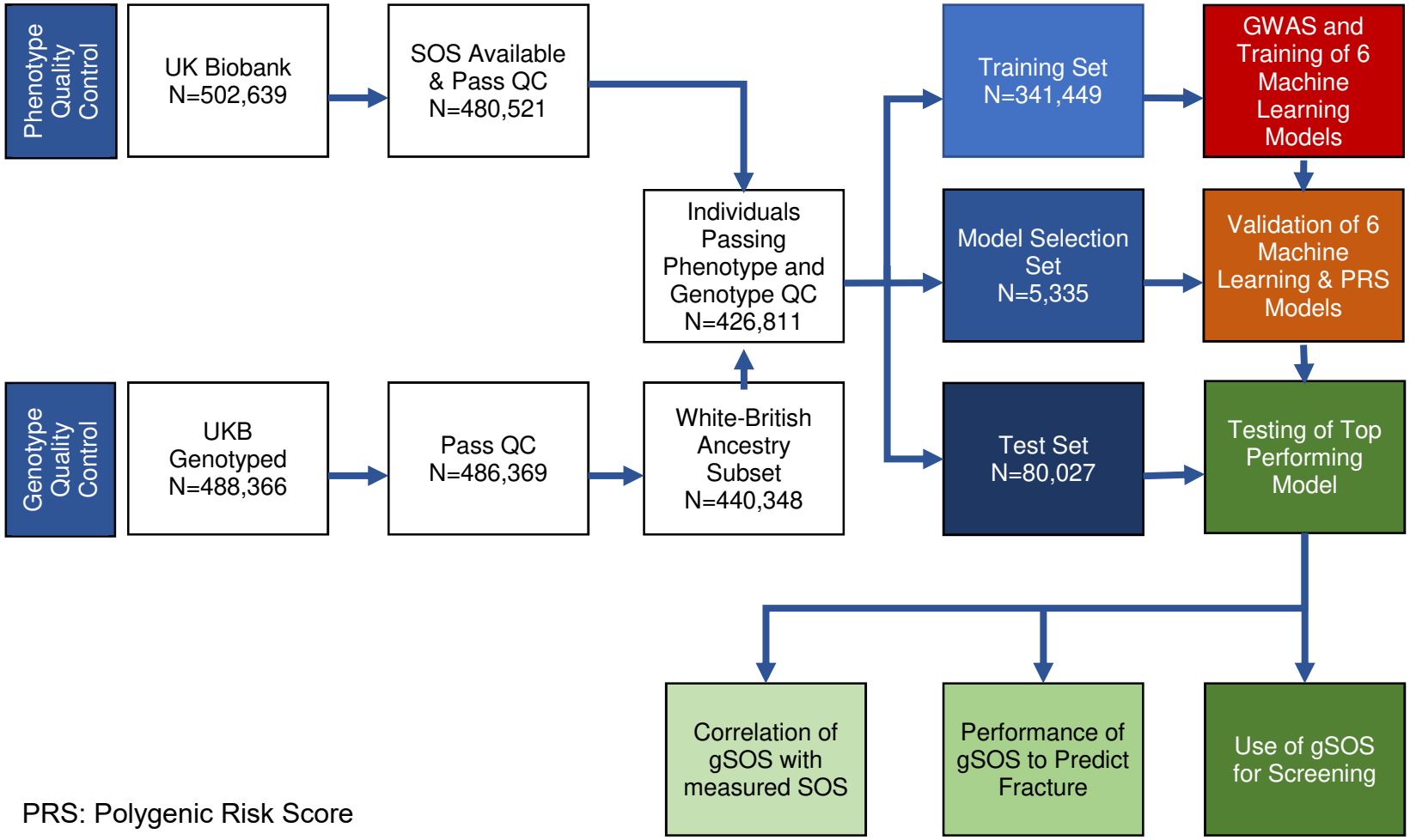
Acknowledgements
This research has been conducted using the UK Biobank Resource under project number 24268. We appreciate the benevolence of UK Biobank volunteers.

References
1 Shendure J, Balasubramanian S, Church GM, *et al.* DNA sequencing at 40: past, present and future. *Nature* 2017; **550**: 345–53.
2 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. Nat. Rev. Genet. 2012; **13**: 395–405.
3 Grzymski JJ, Coppes MJ, Metcalf J, *et al.* The Healthy Nevada Project : rapid recruitment for population health study 2215 Raggio Parkway 2215 Raggio Parkway. 2018; : 1–24.
4 Zheng H, Forgetta V, Hsu Y-H, *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* 2015; **526**: 112–7.
5 Kemp JP, Morris JA, Medina-Gomez C, *et al.* Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat Genet* 2017; **49**: 1468–75.
6 Brent Richards J, Zheng H-FH-F, Spector TDTDTD, *et al.* Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nat Rev Genet* 2012; **13**: 672–672.
7 Howard GM, Nguyen T V., Harris M, Kelly PJ, Eisman JA. Genetic and Environmental Contributions to the Association Between Quantitative Ultrasound and Bone Mineral Density Measurements: A Twin Study. *J Bone Miner Res* 1998; **13**: 1318–27.
8 Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis. *Am J Med* 1993; **94**: 646–50.
9 Papaioannou A, Morin S, Cheung AM, *et al.* 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. *Cmaj* 2010; **182**: 1864–73.
10 Compston J, Cooper A, Cooper C, *et al.* UK clinical guideline for the prevention and treatment of osteoporosis. *Arch Osteoporos* 2017; **12**. DOI:10.1007/s11657-017-0324-5.
11 Cosman F, Lindsay R, LeBoff MS, Jan de Beur S, Tanner B. National Osteoporosis Foundation. Clinician's Guide to Prevention and Treatment of Osteoporosis. Washington, DC: National Osteoporosis Foundation; 2014. *Natl Osteoporos Found* 2014; **1**: 55.
12 Pfister AK, Welch CW, Emmett M, Nordin C. Screening for osteoporosis: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med* 2011; **155**: 276; author reply 276-7.
13 Kanis JA on behalf of the World Health Organization Scientific Group. Assessment of osteoporosis at the primary health care level: Technical Report. 2008. http://www.shef.ac.uk/FRAX/index.htm.
14 Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAXtrade mark and the assessment of fracture probability in men and woman from the UK. *Osteoporos Int* 2008; **19**: 385–97.
15 Shepstone L, Lenaghan E, Cooper C, *et al.* Screening in the community to reduce fractures in older women (SCOOP): A randomised controlled trial. *Lancet* 2017; : 741–7.
16 Meigs JB, Shrader P, Sullivan LM, *et al.* Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med* 2008; **359**: 2208–19.
17 Estrada K, Styrkarsdottir U, Evangelou E, *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet* 2012; **44**: 491–501.
18 Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum*

Mol Genet 2009; **18**: 3525–31.
19 Khera A V., Chaffin M, Aragam KG, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018; : 1.
20 Inouye M, Abraham G, Nelson CP, *et al.* Genomic risk prediction of coronary artery disease in nearly 500,000 adults: implications for early screening and primary prevention. *bioRxiv* 2018; : 250712.
21 Thériault S, Lali R, Chong M, Velianou JL, Natarajan MK, Paré G. Polygenic Contribution in Individuals With Early-Onset Coronary Artery DiseaseClinical Perspective. *Circ Genom Precis Med* 2018; **11**: e001849.
22 Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* 2017. DOI:10.1038/nrg.2017.101.
23 Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; : 1–11.
24 McCloskey E V., Kanis JA, Odén A, *et al.* Predictive ability of heel quantitative ultrasound for incident fractures: an individual-level meta-analysis. *Osteoporos Int* 2015; **26**: 1979–87.
25 Tibshirani R. Regression Selection and Shrinkage via the Lasso. J. R. Stat. Soc. B. 1996; **58**: 267–88.
26 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* 2015; **12**: e1001779.
27 Bycroft C, Freeman C, Petkova D, *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 2017; : 166298.
28 McCarthy S, Das S, Kretzschmar W, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; published online Aug 22. DOI:10.1038/ng.3643.
29 Loh P-RR, Tucker G, Bulik-Sullivan BK, *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015; **47**: 284–90.
30 Kanis JA, Oden A, Johansson H, Borgström F, Ström O, McCloskey E. FRAX® and its applications to clinical practice. Bone. 2009; **44**: 734–43.
31 Mackey DC, Lui LY, Cawthon PM, *et al.* High-trauma fractures and low bone mineral density in older women and men. *JAMA* 2007; **298**: 2381–8.
32 Sanders KM, Pasco JA, Ugoni AM, *et al.* The exclusion of high trauma fractures may underestimate the prevalence of bone fragility fractures in the community: the Geelong Osteoporosis Study. *J Bone Min Res* 1998; **13**: 1337–42.
33 Kanis JA, Oden A, Johnell O, *et al.* The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 2007; **18**: 1033–46.
34 Papaioannou A, Morin S, Cheung AM, *et al.* Clinical Practice Guidelines for the Diagnosis and Management of Osteoporosis in Canada : Background and Technical Report. *Clin Pract Guidel Osteoporos Backgr Tech Rep* 2010; : 1–89.
35 Walter K, Min JL, Huang J, *et al.* The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**: 82–90.
36 Nayak S, Roberts MS, Greenspan SL. Cost-Effectiveness of Different Screening Strategies for Osteoporosis in Postmenopausal Women. *Ann Intern Med* 2011; **155**: 751.
37 Turner DA, Khioe RFS, Shepstone L, *et al.* The Cost-Effectiveness of Screening in the Community to Reduce Osteoporotic Fractures in Older Women in the UK: Economic Evaluation of the SCOOP Study. *J Bone Miner Res* 2018; **33**: 845–51.
38 Richards JB, Leslie WD, Joseph L, *et al.* Changes to Osteoporosis Prevalence According to Method of Risk Assessment. *J Bone Miner Res* 2006; **22**: 228–34.

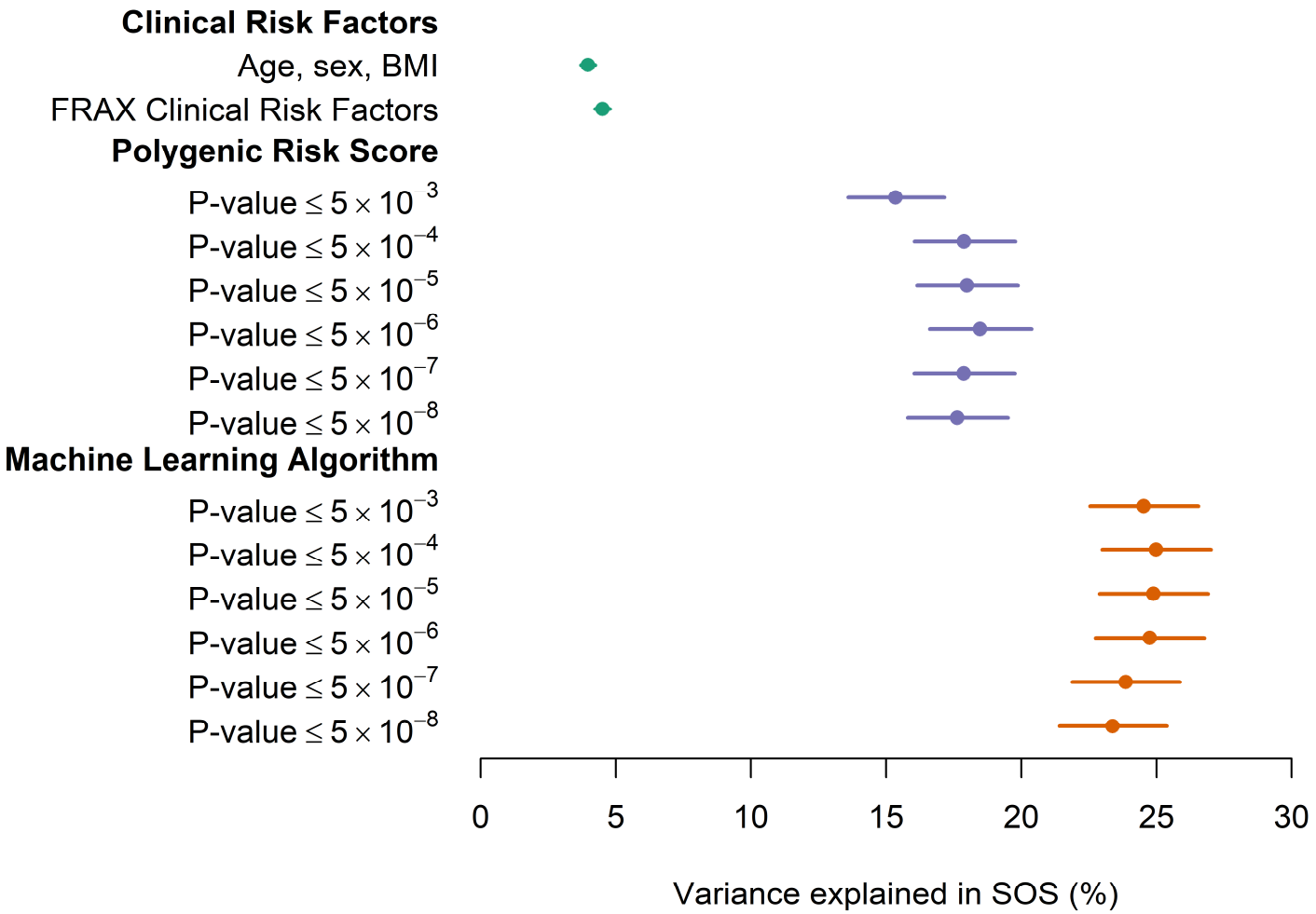
39 Eriksson J, Evans DS, Nielson CM, *et al.* Limited clinical utility of a genetic risk score for the prediction of fracture risk in elderly subjects. *J Bone Miner Res* 2015; **30**: 184–94.
40 Lello L, Avery SG, Tellier L, Vazquez A, Campos G de los, Hsu SDH. Accurate Genomic Prediction Of Human Height. *bioRxiv* 2017; : 1–17.
41 Kim SK. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. *PLoS One* 2018; **13**: 1–20.
42 Allen N, Sudlow C, Downey P, *et al.* UK Biobank: Current status and what it means for epidemiology. *Heal Policy Technol* 2012; **1**: 123–6.

Figure 1. Overall Study Design



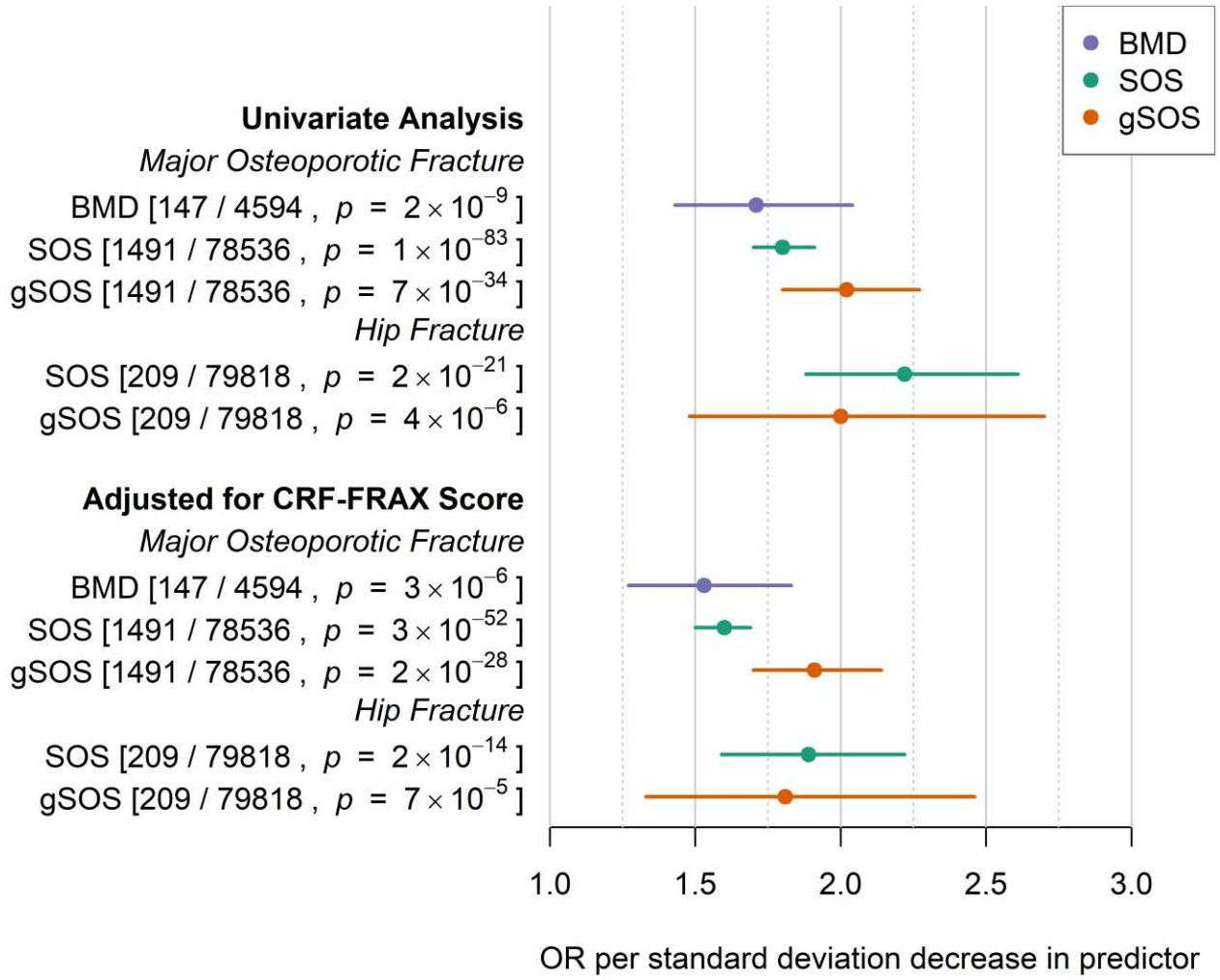
PRS: Polygenic Risk Score

Figure 2. Variance Explained in SOS by Clinical Risk Factors, Polygenic Risk Score and Machine Learning Algorithm in the Model Selection Set



All estimates are from the Model Selection Set and Table S2 provides additional details.

Figure 3. Performance of SOS, gSOS and BMD for the Prediction of Major Osteoporotic and Hip Fractures in the Test Dataset



Sample size of cases / controls are shown in square brackets, followed by P-values

Table 1: Characteristics of Participants by Dataset

	Dataset			
	Training	**Model Selection**	**Test**	**Individuals with DXA BMD (a subset of the Test Dataset)**
Sample Size	341449	5335	80027	4741
SOS	-0.02 (0.99)	-0.03 (0.97)	-0.02 (0.99)	0.13 (1.01)
gSOS	NA	-0.02 (0.45)	-0.02 (0.45)	0.00 (0.44)
NHANES T-Score*	NA	NA	-0.64 (1.04)	-0.64 (1.04)
Age	56.83 (8.00)	56.59 (8.12)	56.77 (8.03)	55.81 (7.56)
Women	186,569 (55%)	2,863 (54%)	43,732 (55%)	2,489 (0.52%)
BMI	27.34 (4.70)	27.34 (4.65)	27.34 (4.70)	26.81 (4.35)
Smoker	27,181 (7.96%)	397 (7.44%)	6,909 (8.63%)	966 (20%)
Rheumatoid Arthritis	3,312 (0.97%)	41 (0.77%)	758 (0.95%)	28 (0.59%)
Corticosteroid Users	3,330 (0.98%)	51 (0.77%)	839 (1.05%)	70 (1.48%)
Secondary Osteoporosis	14,541 (4.26%)	215 (4.03%)	3,354 (4.18%)	192 (4.05%)
Prevalent Fracture	34,917 (10.2%)	549 (10.3%)	8,244 (10.3%)	438 (8.46%)
Incident Hip Fracture	994 (0.29%)	12 (0.22%)	209 (0.26%)	1 (0.02%)
Incident Major Osteoporotic Fracture	6,307 (1.85%)	84 (1.57%	1,491 (1.86%)	157 (3.0%)

Data are presented as mean and (standard deviation), or counts and percentage for binary traits)
SOS is standardized to have a mean of zero and a standard deviation of one.
BMI: Body Mass Index

Table 2. Correlation of gSOS with FRAX Clinical Risk Factors

Variable	Measured SOS			gSOS	
	r	P-value		r	P-value
Age	-0.12	$<10^{-4}$		0.006	0.11
Sex	-0.15	$<10^{-4}$		-0.003	0.41
BMI	0.035	$<10^{-4}$		-0.002	0.63
Smoking	-0.05	$<10^{-4}$		0.001	0.74
Corticosteroid Use	-0.016	$<10^{-4}$		0.0005	0.89
Rheumatoid Arthritis	-0.04	$<10^{-4}$		-0.006	0.09
Secondary Causes of Osteoporosis	-0.074	$<10^{-4}$		-0.005	0.16

Age is in years, BMI kg/m2, women are labeled as "1" and all other variables were labeled as "1" if risk factor was present.

Table 3. Prediction of Incident Fractures by FRAX Scores

	Major Osteoporotic Fracture				Hip Fractures		
Predictor	**Odds per 1 log increase in predictor (95% CI)**	**AUC (95% CI)**	**Comparison of AUC with gSOS AUC[a]**		**Odds per 1 log increase in predictor (95% CI)**	**AUC (95% CI)**	**Comparison of AUC with gSOS AUC[a]**
CRF-FRAX	1.68 (1.60-1.76)	65% (64%-66%)	gSOS improves AUC ($P = 3x10^{-14}$)		2.95 (2.54-3.42)	77% (74%-80%)	gSOS improves AUC ($P = 4x10^{-3}$)
BMD-FRAX	1.69 (1.47-1.96)	66% (62%-71%)	gSOS equivalent to BMD (P = 0.47)		N/A	N/A	N/A
SOS-FRAX	1.91 (1.82-2.01)	68% (67%-70%)	SOS improves AUC ($P = 7x10^{-9}$)		2.67 (2.36-3.02)	79% (76%-82%)	SOS equivalent to gSOS (P = 0.09)
gSOS-FRAX	1.77 (1.69-1.86)	67% (65%-68%)	-		2.94 (2.55-3.39)	78% (75%-81%)	-

N/A: Too few cass of hip fractures amongst 4741 individuals with BMD measures
CRF-FRAX: Clinical Risk Factors FRAX Score
BMD-FRAX: Clinical Risk Factors plus BMD FRAX Score
SOS-FRAX: Clinical Risk Factors plus SOS FRAX Score
gSOS-FRAX: Clinical Risk Factors plus gSOS FRAX Score
AUC: Area under the receiver operator cure
a) Comparison with gSOS compares the AUC for gSOS with each of the other predictors

Table 4. Effect of gSOS Pre-Screening on NOGG Screening Tests and Outcomes.

		Screening: Standard of Care	Screening After Introduction of gSOS Pre-Screen
Number of Participants	**Number of Participants with Sufficient Data Available to Test Efficacy of Screening**	4741	
	Number of Participants Eligible for Screening	2455	
Clinical Visits, BMD Tests and Outcomes	**Number of Participants Requiring a Clinical Visit**	2455	1273
	Number of Participants Requiring a BMD Test	1013	473
	Number of Participants Reassured After Screening	2223	2238
	Number of Participants Recommended for Treatment	232	217
Performance	**Sensitivity For Correct Treatment Allocation[a]**	99.1%	94.7%
	Specificity For Correct Treatment Allocation	99.6%	99.9%
	Positive Predictive Value	96.5%	98.6%
	Negative Predictive Value	99.9%	99.5%

Table S4 provides the outcomes of each participant
Figure S4 and S6 provides a flow chart demonstrating the application of NOGG guidelines and gSOS Pre-Screening
Only participants in the test dataset with BMD measures were included in this analysis to enable an evaluation of the performance of guidelines to correctly identify who should receive therapy, compared to the situation where all individuals had BMD measured
a. "Correct treatment allocation" is defined as assignment to reassurance or treatment by BMD-FRAX.