Article:
Cogurcu, Y.E., Bartolomei, M., James, B.A. et al. (2026) Evaluating digital twin visualisations for safety and trust in robot-assisted dressing. ACM Transactions on Human-Robot Interaction, 15 (4). pp. 1-33. ISSN: 2573-9522

https://doi.org/10.1145/3805805

ESHG

www.nature.com/ejhg

ARTICLE OPEN

Check for updates

Reduced penetrance of *COL1A1/2* pathogenic variants linked with osteogenesis imperfecta: analysis of a large population cohort

Alistair T. Pagnamenta[1], James Fasham[1], Robin N. Beaumont[1], Duncan Baker[2], Sylvia Keigwin[2], Tim Hall[1], Emma L. Baple[1,3], Meena Balasubramanian[4,5], Leigh Jackson[1] and Caroline F. Wright[1]✉

Osteogenesis imperfecta (OI) is under consideration for inclusion in several genomic newborn screening initiatives, but its penetrance in clinically-unselected populations is currently unknown. It is an exemplar condition for evaluating penetrance in adult cohorts due to its relatively low mortality, variable expressivity and link to several large genes. Using genome sequencing data from ~500,000 adults in UK Biobank, we curated a set of rare pathogenic/likely pathogenic (P/LP) variants in *COL1A1, COL1A2* and *IFITM5* using annotations from gnomAD, ClinVar and SpliceAI. Analysis of summed read-count data from genome and exome sequencing led to exclusion of 16 mosaic variants with consistently low allelic balance of 3.4–35.9%. We identified 61 likely constitutive heterozygous P/LP variants in *COL1A1* and *COL1A2* (29 loss-of-function or splice variants and 32 missense) in 115 participants, with a mean age at recruitment of 55.1 years; no P/LP variants were identified in *IFITM5*. Phenotypes were assessed using ICD-10 codes, self-reports and heel bone mineral density (BMD). Overall disease penetrance was lower than anticipated: 40.7% for *COL1A1* and 21.3% for *COL1A2*, potentially due to depletion of severe early-onset disease. When considering only truncating variants in *COL1A1*, disease penetrance increased to 73.1%, and 90% of individuals had reduced levels of circulating COL1A1 protein. For *COL1A2*, BMD data supported the low penetrance, whereas for *COL1A1*, data suggested the possibility of a subclinical phenotype. Overall, for P/LP missense variants (including those altering Gly-Xaa-Yaa repeats), the low penetrance observed suggests reliance on current ClinVar assertions to support pathogenicity may overstate OI risk in population screening.

European Journal of Human Genetics; https://doi.org/10.1038/s41431-026-02118-6

INTRODUCTION

Newborn screening (NBS) enables early detection of selected genetic/metabolic disorders to improve clinical management. Well-defined criteria are used to determine which conditions are included [1]. For genetic conditions, screening is limited to severe, early-onset diseases with high penetrance, i.e. where a pathogenic/likely pathogenic (P/LP) variant has >80–90% chance of causing disease [2]. Penetrance estimates are typically derived from clinical cohorts and may suffer from ascertainment bias [3]. A recent study integrating electronic health records with predicted loss-of-function (pLOF) variants across genes confirmed to cause disease via genetic haploinsufficiency, indicates that incomplete penetrance is not uncommon [4]. Therefore, to avoid over-reporting, systematic re-evaluation using large population cohorts is required.

Many countries are assessing the potential benefits of using whole-genome sequencing (WGS) to expand NBS to include increased numbers of genes/variants [5]. One such initiative is the UK's Generation Study (GS) [6], which commenced recruitment in 2024 and currently reports P/LP variants in >450 genes. Despite extensive guidance, consensus on gene inclusion remains limited. A comparison of 27 international genomic NBS (gNBS) studies investigated >4000 genes, but only 18 were included in all initiatives [5].

Osteogenesis imperfecta (OI) is a heterogeneous bone dysplasia, characterized by low bone mass and increased fragility [7, 8]. Affected individuals are at increased risk of long-bone and vertebral compression fractures. Variable deformities of the long-bones/ribs/spine may be observed from birth, often with substantial growth deficiency. Other features include blue sclerae and hearing loss, and the prevalence is estimated at 0.3–1.5 cases per 10,000 individuals [7, 9, 10]. Of 36 genes conclusively linked to OI in PanelApp [11, 12] and PanelApp Australia's equivalent panel [13], and included in the NHS Genomic Medicine Service Test Directory [14], only three are tested in the GS. These include *COL1A1* (MIM: 120150) and *COL1A2* (MIM: 120160), which together represent the most common cause of OI (types I-IV) [7, 9, 15], and *IFITM5* (MIM: 614757), linked to a severe OI subtype caused by a much smaller repertoire of activating variants [16]. Although some of the autosomal recessive OI subtypes can also result in severe

[1]Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, UK. [2]Connective Tissue Disorders Service, Sheffield Diagnostic Genetics Service, Sheffield Children's Hospital, Sheffield, UK. [3]Department of Clinical Genetics, Royal Devon University Hospital, Exeter, UK. [4]Division of Clinical Medicine, University of Sheffield, Sheffield, UK. [5]Highly Specialised Severe, Complex and Atypical OI Service, Sheffield Children's NHS Foundation Trust, Sheffield, UK.
✉email: caroline.wright@exeter.ac.uk

Received: 30 January 2026 Revised: 7 April 2026 Accepted: 15 April 2026
Published online: 12 May 2026

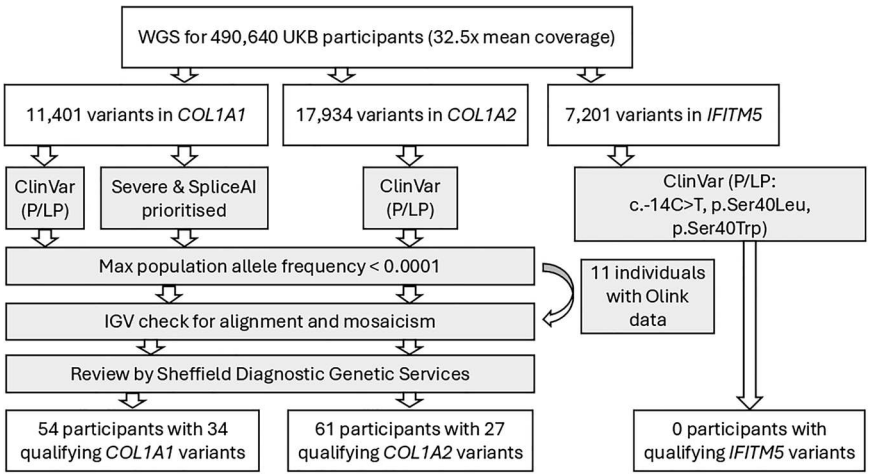

Fig. 1 Schematic flow-diagram summarising variant filtering steps in study. Numbers of qualifying variants and numbers of UK Biobank participants harbouring this set of constitutive (likely) pathogenic variants are shown at the bottom. Intermediate filtering steps are highlighted in grey.

clinical presentations (e.g. types VII and VIII, caused by pathogenic variants in *CRTAP* and *P3H1*, respectively), these only represent a small minority of cases in the UK [17], are included in relatively few gNBS panels internationally [5], and no gene-specific therapies are currently established.

COL1A1 and *COL1A2* are closely related paralogs that encode α1/2 collagen chains. These chains form a triple-helix, composed of two α1 and one α2 molecules. Due to its tightly-coiled nature [18], every third residue is glycine. Variants that disrupt this Gly-Xaa-Yaa pattern substantially disrupt the triple-helix and result in phenotypes, through a dominant-negative mechanism [7]. Of the four distinct subtypes of OI caused by *COL1A1/2* variants, types II (often with respiratory insufficiency) and III (a 'progressive deforming' form) are the most severe [19]. The milder OI type I is typically caused by *COL1A1* pLOF variants [15]. Predicting clinical outcome is not straightforward, particularly for previously unreported variants. Data from clinical cohorts suggests that although expressivity varies widely, disease penetrance in individuals carrying a heterozygous P/LP variant in *COL1A1* or *COL1A2* is near complete [8].

Management of newborns with suspected OI involves careful handling from birth, with avoidance of manual hip examination. Treatment may also include bisphosphonate, surgical or physiotherapy interventions and ventilatory support. Early adoption of such interventions, following molecular confirmation, is associated with reductions in early morbidity and mortality [9]. However, the potential benefits of such therapies must be weighed against the potential harms caused by the reporting of false positive results, which can include side-effects of unnecessary treatment and psychological distress. Given the variable expressivity, the value of early genetic testing is an area of active debate. Some of these children may never present with a fracture in their lifetime and therefore the implications for appropriate follow-up remain unclear. In this study, we evaluated the prevalence and penetrance of P/LP variants in *COL1A1, COL1A2* and *IFITM5* across half-a-million UK adults. Following integration with a rich set of clinical, molecular and physiological phenotypes, we assessed gene-level population penetrance in order to evaluate the likely efficacy of including these genes in gNBS.

MATERIALS AND METHODS

WGS and variant annotation

UK Biobank (UKB) is a population-based study involving ~500,000 participants from across the UK, recruited between 2006 and 2010, aged 40–69 [20]. In addition to the collection of comprehensive demographic and health-related measures, exome ($N = 454,787$) and WGS ($N = 490,640$) were performed [21, 22]. Variant curation primarily focused on WGS data, generated using 150 bp paired-end reads on the NovaSeq6000 (Illumina). Mapping to GRCh38 and variant calling used the DRAGEN pipeline. *COL1A1*, *COL1A2* and *IFITM5* variants were annotated using Ensembl VEP (v110) [23], based on the respective MANE Select transcripts NM_000088.4, NM_000089.4 and NM_001025295.3.

Filtering and QC of variants

Different variant prioritisation approaches were taken for each of the three candidate genes (Fig. 1), as detailed below.

- For *COL1A1*, the ClinGen Dosage Sensitivity Curation Working Group verified that sufficient evidence exists to support haploinsufficiency as a disease mechanism for OI (CCID:006903), which includes pathogenic deletions/frameshifts/start-loss/stop-gains [24–26]. All pLOF variants were therefore prioritised, regardless of whether they had been reported previously in ClinVar [27], or elsewhere. Candidate splice variants were assessed using SpliceAI-visual [28], with absolute as well as delta SpliceAI scores; four exemplar variants in *COL1A1* (Fig. S1, Note S1) demonstrate how these can assist interpretation. Missense variants are also known to cause OI and these were prioritised primarily using P/LP ClinVar assertions, with emphasis on those impacting the Gly-Xaa-Yaa repeat structure.
- For *COL1A2*, ClinGen assessment indicates that there is no evidence for haploinsufficiency as a disease mechanism (CCID:006904) and most OI-associated variant in ClinVar are missense. For this gene, prioritisation was primarily based on ClinVar pathogenicity assessments.
- For *IFITM5*, prioritisation was straightforward, as there are only three activating variants known to be P/LP. These include a highly recurrent 5′-UTR variant (c.-14C>T) and two missense substitutions involving codon 40.

All variants were filtered using a population allele-frequency of <0.0001 (gnomAD v3) and a UKB allele-count of <10. Sequencing reads were visually inspected for all variants using IGV (v2.6.6) [29] and any with poor mapping-quality, low read-depth, allelic imbalance, or strand bias were excluded. Analysis of potential somatic mosaicism involved summing allelic read-counts from both exome and WGS datasets. Cumulative binomial tests were performed to evaluate whether the variant allelic fraction (VAF) differed significantly from the 50% expected for constitutive germline variants. All qualifying variants were then independently assessed by the Sheffield Diagnostic Genetic Services [30], which provides the UK national targeted testing services for OI. This last step was performed blind to any phenotype information from UKB.

Phenotype analysis

Once the set of high-quality qualifying variants had been finalised, individuals carrying these variants were identified using an in-house app in

the DNANexus research user environment. The R package ukbrapR was used to identify UKB participants with the ICD-10 code Q78.0 (Osteogenesis Imperfecta). Self-reported illnesses from the verbal questionnaire were reviewed for evidence of any bone-related phenotypes. For further granularity of phenotype data, another custom R script was used to extract all available medical records in a strict chronological order. This data included ICD-9/10 [31], GP records and hospital episode statistics.

Proteomic and bone mineral density (BMD) assessments

Proteomics data from blood plasma was generated in UKB participants using the Olink technology (Note S2) [32]. Olink results are available for 2923 circulating proteins (including COL1A1) as "Normalized Protein eXpression" (NPX) scores, a log2-normalised quantification of protein levels, where a 1-unit change reflects a two-fold difference in abundance. To assess reproducibility, data from the first batch ('instance 0') was compared to data from subsequent batches ('instances 2 and 3'). Heel BMD (h-BMD) measurements were also available for 278,754 UKB participants, based on the quantitative ultrasound index through the calcaneus. Standard Polygenic Risk Score (PRS) for estimated BMD t-score was based on external GWAS data [33]. All data was downloaded using the Table exporter (DNAnexus) and analysed with RStudio (v4.2).

RESULTS

Manual variant curation and exclusion of somatic mosaic variants yields a high-quality set of variants in UKB expected to cause OI types I-IV

Across all UKB participants, there were a total of 36,536 SNV/indels with annotations on the MANE Select transcripts for *COL1A1*, *COL1A2* or *IFITM5*. Of these, 74 met our initial variant inclusion criteria.

Across *COL1A1/2*, 15 SNVs and 1 indel were detected at significantly reduced VAF, suggesting somatic mosaicism (Table 1). Ten of these were in *COL1A1* and VAF ranged from 3.4% for c.2299G>A, p.(Gly767Ser) to 35.9% for c.1598G>A, p.(Gly533Asp). The remaining six in *COL1A2* ranged from 11.6% for c.2314G>A, p.(Gly772Ser) to 32.1% for c.767G>T, p.(Gly256Val). Four variants were observed in both mosaic and non-mosaic states; for these only the individuals with constitutive variants were retained (Tables S1, S2). Allelic ratios obtained from exome and WGS showed a strong correlation ($R^2 = 0.60$; Fig. S2), suggesting these are true mosaic variants. Alignments for representative examples of a mosaic indel (c.2766_2769del) and SNV (c.1A>G) in *COL1A1* are shown (Fig. S3A-B). Review of sequence reads also identified co-occurring variants. Two examples of this were c.1602del and c.1604del in *COL1A1*, observed in trans (Fig. S3C), whilst c.1A>C and c.2_5del were *in cis* (Fig. S3D). In each case, only the better supported variant was retained (Table S1). Further literature review led to additional exclusions (Note S3).

Following manual curation, *COL1A1* had 34 qualifying heterozygous variants in 54 UKB participants (Fig. 1, Table S1), while *COL1A2* contained 27 qualifying heterozygous variants in 61 participants (Fig. 1, Table S2). No P/LP CNVs in *COL1A1/2* were identified (Note S4). Of these 61 high-quality variants, 47 were singletons and the most frequent (NM_000089.4:c.3196G>A and NM_000089.4:c.2701G>A) occurred in nine individuals each. There were 29 pLOF/splice variants and 32 missense substitutions, 29 of which altered a glycine in the Gly-Xaa-Yaa repeat regions. Of the 61 variants, 45 were SNVs and 16 were indels. CADD scores ranged from 16 (NM_000088.4:c.1354-12G>A) to 49 (NM_000088.4:c.1981C>T; p.Gln661Ter). Most qualifying variants (53/61) were listed in ClinVar at the time of assessment (31 P, 10 LP/P and 10 LP), with two of the retained variants currently listed as 'conflicting'. For instance NM_000089.4:c.1423G>A (p.Gly475Ser) was LP/P in a previous version (VCV001398959.6), but a single VUS assessment in the current version (VCV001398959.8) moves the overall classification to 'conflicting'. Given the 3 P/LP classifications and the fact it is expected to disrupt the triple helical domain, this variant was retained. For *COL1A1*, start-loss is a known disease mechanism [24] and so NM_000088.4:c.2_5del was also retained. Overall, variants were widely distributed across both genes (Tables S1-2, Fig. S4A,B).

Despite minimal depletion of OI in UKB, penetrance of P/LP variants in COL1A1/2 was lower than in clinical cohorts

Extrapolation from the reported prevalence of OI [7, 9, 10] predicts between 15 and 75 individuals with OI amongst ~500k individuals within UKB. Based on ICD-10 code Q78.0, there were 76 individuals in UKB with OI, a value close to the upper-bound of the expected range, consistent with minimal depletion due to "healthy volunteer" selection bias [34]. Although the most severe (childhood-lethal) forms of OI would not be expected to be present in an adult cohort, this prevalence suggests that clinical data in UKB is likely to be largely complete for the OI phenotype. Nevertheless, to increase confidence we also scrutinised all available electronic health records (including hospital episode statistics and GP records) chronologically, searching for the presence of multiple fracture events that could not be explained by environmental factors or other events captured in the records. Individuals with fractures in close temporal proximity to V43.5/V18.4 or W51 codes (driver/cyclist injured in traffic accident, accidental striking against or bumped into by another person) were not considered affected.

Using all relevant health records to assess OI phenotypes, the estimated penetrance across UKB for adult carriers of P/LP variants in *COL1A1* was 40.7% (22/54), with most positive cases derived from ICD-10 coding (Fig. 2A). Restricting the set of variants to only pLOF variants in *COL1A1*, penetrance estimates increased to 73.1% (19/26), whilst penetrance for missense *COL1A1* variants was only 10.7% (3/28) (Fig. 2B, C). In contrast, estimated penetrance for carriers of P/LP variants in *COL1A2* was 21.3% (13/61), with just five identified through ICD-10 codes (Fig. 2D). These results are substantially lower than penetrance estimates based on clinical cohorts, where penetrance for heterozygous *COL1A1* or *COL1A2* P/LP variants has been estimated at close to 100% [8]. Most ICD-10 diagnoses in carrier individuals occurred between the ages of 40–60 (Fig. 2E).

Proteomic data aids variant curation and supports functional effect of pLOF variants in COL1A1

There were 52,155 UKB participants with Olink data for COL1A1. The 25th centile for NPX score was at −0.1847 whilst the 1st centile was at −0.645, with just 522/52,155 results lying below the latter threshold. Of the participants with qualifying variants in *COL1A1*, 11 had Olink data available and by chance, these all involved pLOF variants. Reassuringly, the *COL1A1* NPX scores for this set of individuals mostly clustered around -1, consistent with heterozygous pLOF variants resulting in a 50% drop in protein levels (Fig. 3). Interestingly, a high NPX score for a known pathogenic start-loss variant in one of these individuals helped support its exclusion due to likely somatic mosaicism (Table 1, Fig. S3B). In the remaining 10 individuals with pLOF variants in *COL1A1* and Olink results, levels of α1 collagen protein were <1st centile in nine, of whom eight also had the ICD-10 code for OI (Q78.0).

A reciprocal analysis was performed starting with all 76 individuals in UKB with ICD-10-defined OI. Of these, only 13 had Olink data available and of those, nine had an NPX score <1st centile. From these nine individuals, eight harboured a qualifying *COL1A1* variant detected by WGS (i.e. the same variants/individuals mentioned above). The only additional individual in UKB with Q78.0 and a reduced level of α1 collagen but without a *COL1A1* variant in the WGS data was subsequently found to carry a c.3548del, p.(Pro1183LeufsTer56) variant, detected by exome sequencing (WGS unavailable for this participant). These results highlight that the combination of a clinical OI diagnosis and an Olink-confirmed reduction of α1 collagen is highly specific for the presence of a *COL1A1* pLOF variant. For ~1000 individuals with Olink data replicates, modest correlations between NPX scores

Table 1. Summary of (likely) pathogenic variants suspected to be mosaic and removed from penetrance calculations.

HGVS (cDNA)	HGVS (protein)	alt genome	Ref genome	VAF genome	Alt exome	Ref exome	Vaf exome	Alt combined	Ref combined	Vaf combined	*P*-value
COL1A1 (ENST00000225964.10)											
c.1598G>A	p.(Gly533Asp)	7	22	31.8%	30	81	37.0%	37	103	35.9%	0.0028
c.769G>A	p.(Gly257Arg)	11	28	39.3%	21	65	32.3%	32	93	34.4%	0.0017
c.1921G>A	p.(Gly641 Arg)	11	40	27.5%	17	42	40.5%	28	82	34.1%	0.0027
c.1A>G†	p.(Met1?)	12	57	21.1%	17	48	35.4%	29	105	27.6%	3E-06
c.2902G>T	p.(Gly968Ter)	5	27	18.5%	7	48	14.6%	12	75	16.0%	8E-10
c.904-2A>C	—	4	35	11.4%	10	54	18.5%	14	89	15.7%	2E-11
c.3082G>A	p.(Gly1028Ser)	6	46	13.0%	10	59	16.9%	16	105	15.2%	9E-14
c.2766_2769del	p.(Gly923ProfsTer184)	6	36	16.7%	8	76	10.5%	14	112	12.5%	5E-17
c.369+1G>A	—	4	20	20.0%	2	80	2.5%	6	100	6.0%	1E-21
c.2299G>A	p.(Gly767Ser)	3	25	12.0%	2	122	1.6%	5	147	3.4%	3E-36
COL1A2 (ENST00000297268.11)											
c.767G>T	p.(Gly256Val)	13	37	35.1%	14	47	29.8%	27	84	32.1%	0.0007
c.838G>A	p.(Gly280Ser)	13	28	46.4%	13	53	24.5%	26	81	32.1%	0.0008
c.70+717A>G	—	11	37	29.7%	0	0	NA	11	37	29.7%	0.0100
c.704G>A	p.(Gly235Asp)	9	34	26.5%	12	53	22.6%	21	87	24.1%	7E-07
c.767G>A	p.(Gly256Asp)	10	48	20.8%	12	49	24.5%	22	97	22.7%	3E-08
c.2314G>A	p.(Gly772Ser)	3	25	12.0%	5	44	11.4%	8	69	11.6%	2E-11

Alt, number of reads with variant. Ref, number of reads at variant position. VAF, variant allelic fraction. *P*-values were calculated based on the binomial distribution (one-tailed). Protein coordinates are based on ENSP00000225964.6 (*COL1A1*) and ENSP00000297268.6 (*COL1A2*). For each gene, variants are presented in order of decreasing VAF. †, mosaicism is supported by Olink data which showed the COL1A1 protein levels were not reduced (Fig. 3).

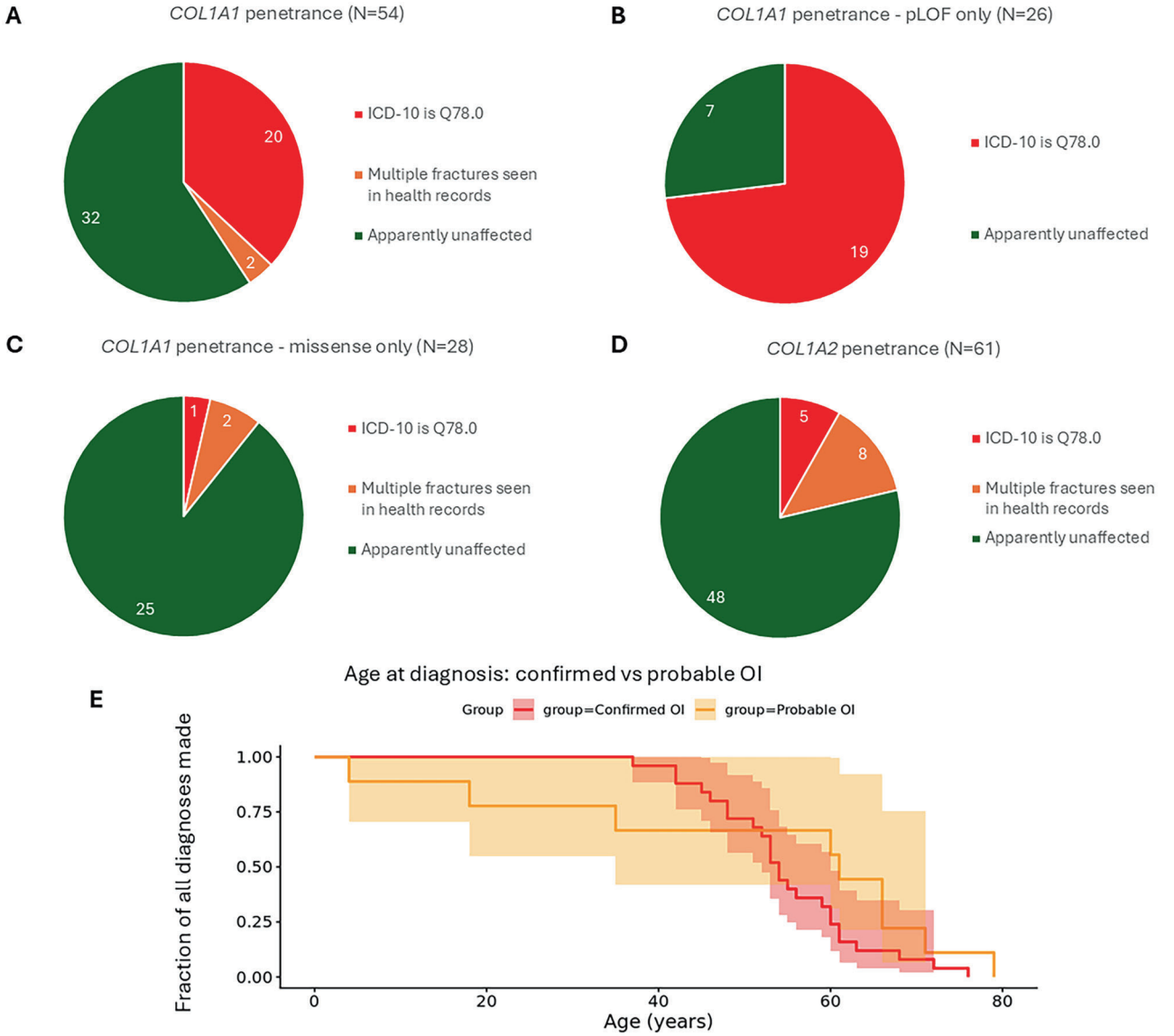

Fig. 2 Incomplete penetrance and age at diagnosis for *COL1A1* and *COL1A2* variant carriers. A Penetrance results for *COL1A1*. The red segment in the pie chart represents individuals with a qualifying variant and where ICD-10 codes include Q78.0. Orange segments denote variant carriers where there is not Q78.0 but where the individual has had multiple fractures. Green segments denote individuals with a variant but where that individual has neither the Q78.0 code nor evidence of multiple fractures. Penetrance estimates were then recalculated using *COL1A1* variants subdivided into (**B**) those that are pLOF (including splice variants) and **C** those which are missense. **D** Penetrance results for carriers of qualifying *COL1A2* variants. **E** Plot showing the approximate ages at which the ICD-10 code Q78.0 was assigned for the 25 variant carriers (range 37-76, shown in red), compared to the 9 individuals where affected status was inferred, based on presence of multiple fractures (4–79 years, orange). For one individual, the date of second fracture was not available and this individual was omitted from the plot. Shaded bands represent 95% confidence intervals.

were observed, and time between blood draws may influence the consistency of circulating protein levels (Fig. S5).

Correlation with BMD supports low penetrance of missense variants in COL1A2

Except for variants that disrupt the procollagen C-propeptide cleavage site [35, 36], OI-associated *COL1A1/2* variants typically decrease BMD. The degree of change depends on whether the variant affects protein qualitatively or quantitatively. Lumbar spine BMD was available for only 45,522 UKB participants and so we used normalised h-BMD, available for 278,975 (mean of -0.337). For individuals who had both measurements (Fig. S6A), L1-L4 vs h-BMD were correlated ($P < 2.2 \times 10^{-16}$, Fig. S6B). For the 27/54 (50%) individuals with qualifying *COL1A1* variants where h-BMD was available, mean BMD was reduced (–1.315; $P = 7.22 \times 10^{-5}$). The mean h-BMD was also reduced in 38/61 (62%) individuals with *COL1A2* variants where data was available (–0.846; $P = 0.016$).

Variant carriers were then split into clinically-affected and unaffected subcategories. Comparisons of BMD measurements between these groups (Fig. 4) revealed no significant difference between BMD in clinically unaffected individuals with P/LP *COL1A2* variants and UKB participants without P/LP *COL1A2* variants ($P = 0.808$). In contrast, for affected individuals with P/LP *COL1A2* variants, mean BMD values decreased from –0.34 to –1.99 ($P < 0.001$). For *COL1A1*, mean BMD for clinically unaffected carriers (–1.23) was intermediate between non-carriers (–0.34) and affected carriers (–1.81). These changes in BMD in clinically-affected individuals due to a single rare variant were over twice

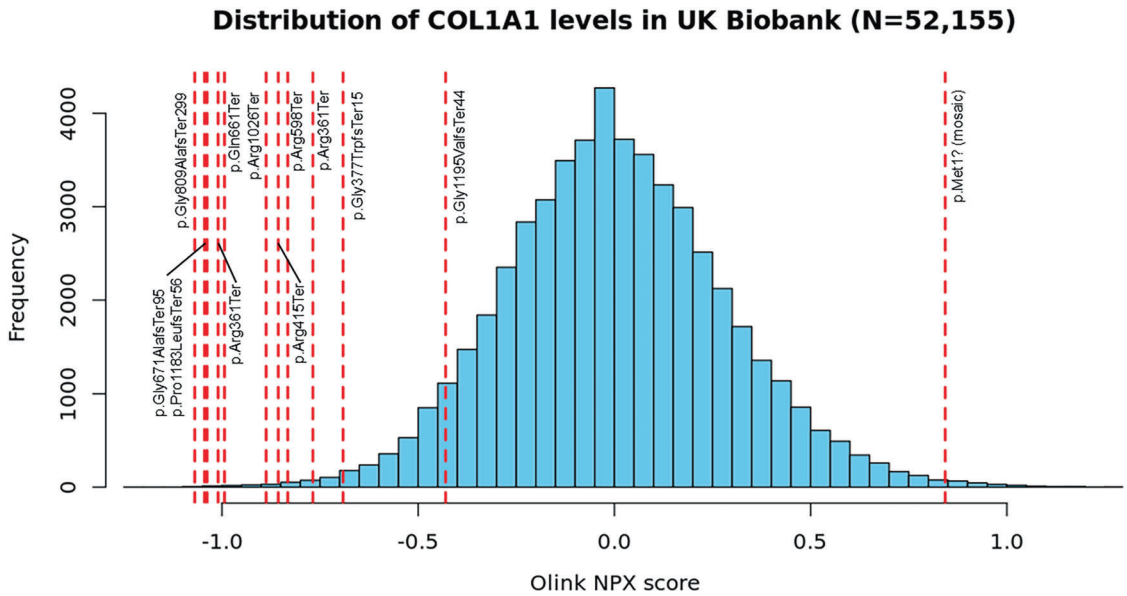

Fig. 3 Proteomic data for *COL1A1* variant carriers. Olink blood proteomics data from UK Biobank supports a deleterious effect of pLOF variants on protein levels. Unfortunately, Olink data was unavailable for carriers of the qualifying missense or splice variants. Histogram was created with RStudio (v4.4) and data for individuals with heterozygous *COL1A1* pLOF variants are labelled. The c.1081C>T, p.(Arg361Ter) variant was detected in two UB biobank participants with Olink data availability. The p.Pro1183LeufsTer56 variant was detected in an individual for whom only exome data was available and so was not included in our penetrance calculations. The c.1A>G (p.Met1?) variant was only present in 27.6% of reads (Fig. S3B, Table 1) and so is likely mosaic, explaining the lack of any decrease in relative protein levels. The p.Gly1195ValfsTer44 variant is the most terminal of all variants shown and thus the protein may retain some stability—this could explain why there is only a partial reduction of protein levels for that sample. All 52,155 datasets are from instance 0 and the mean Normalized Protein eXpression (NPX) value was 0.0038.

those seen when comparing mean BMD results in participants stratified by PRS, where a 0.68 difference was observed between quantiles 1 vs 5 (Fig. S7).

We hypothesized that PRS for BMD might act as a genetic modifier. Such a phenomenon has been shown to occur in other conditions such as developmental delay and hypertrophic cardiomyopathy [37, 38]. However, the BMD PRS did not explain differences in penetrance ($P = 0.304$; Fig. S8).

Lack of P/LP variants in IFITM5 supports extreme rarity and high penetrance of OI type V

No P/LP variants in *IFITM5* were identified in UKB. A different missense variant involving codon 40 was assessed as a VUS, based on ACMG/AMP criteria [39](Note S5).

DISCUSSION

Rare variants in *COL1A1/2* that cause bone fragility are typically detected in around 80–90% of individuals with OI [7, 9, 15]. However, there is ongoing debate about whether these genes should be included in gNBS programs. Of the panels assessed by Minten et al., *COL1A1* was present in 12/27 and *COL1A2* in 13/27 [5]. Furthermore, modelling of the gNBS variant prioritisation strategy from the GS identified *COL1A1/2* as outliers in terms of the number of variants prioritised [40], and so care needs to be taken to limit reporting of variants incorrectly associated with disease, or those with low penetrance. Here, we used genome sequence data from half-a-million adults to show that, for a curated set of P/LP variants in *COL1A1/2*, there is incomplete penetrance (~40% and ~20%, respectively) with respect to a clinical diagnosis of OI and/or the presence of multiple fractures in the available electronic health records. This is in contrast to the literature, as although variable severity is widely reported, penetrance has previously been considered near complete [41].

The OI phenotype resulting from missense variants in *COL1A1*, and *COL1A2* in particular, is typically thought to be more severe than OI due to *COL1A1* haploinsufficiency [7, 41]. This is due to the dominant-negative effect certain missense changes have on the triple-helix structure. Our penetrance estimates were therefore initially counterintuitive, as penetrance for pLOF and splicing variants in *COL1A1* was higher (73.1%) than missense changes in the same gene (10.7%) or in *COL1A2* (21.3%). This likely reflects a combination of misclassified missense variants and depletion of severe childhood-onset disease in UKB, due to healthy volunteer bias [34]. A lack of suitably matched controls in past OI studies and a historical over-reliance on the Gly-Xaa-Yaa rule for variant interpretation we suspect has led to a self-propagating issue of overreporting glycine-altering candidate variants. To combat the potential risks of falsely reporting missense variants detected by gNBS, there will need to be meticulous curation that assesses the most recent literature and integrates clinical information with data from large biobanks, complemented by high-throughput functional testing [42]. Examples of poor or outdated curation in ClinVar are not uncommon (Note S6) and assertions based on single submissions should be treated with caution. Even variant-condition links supported by multiple submissions should be reviewed, especially for genes with multiple phenotype associations (Note S7). Widespread efforts by ClinGen curation teams to flag suspicious or conflicting assertions in ClinVar are ongoing [43, 44].

In contrast to the two α-collagen genes, *IFITM5* was only included in 5/27 of the gNBS panels assessed by Minten et al. [5]. In this study, no *IFITM5* P/LP variants were identified. This lack of pathogenic genotypes across such a large population cohort is consistent with the rarity of OI type V and/or high penetrance, combined with the typically acute childhood phenotypic presentation. The absence of such variants in this adult cohort likely reflects survivor bias. In contrast, we note that five individuals with de novo pathogenic NM_001025295.3:c.-14C>T variant were detected in the 100kGP [45], using a similar methodology. Given this relative absence of P/LP variants in UKB, our results suggest that inclusion of this gene in gNBS would likely achieve a high specificity.

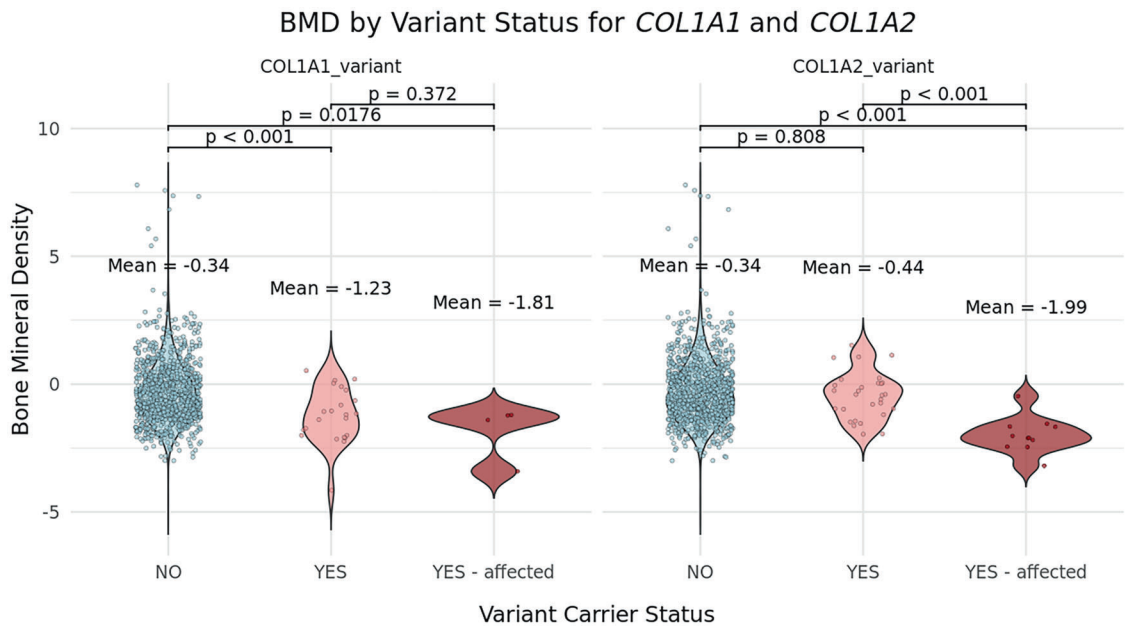

Fig. 4 Bone mineral density (BMD) data for *COL1A1*/*COL1A2* variant carriers. Violin plots show the distribution of heel BMD values across 278,975 individuals from the UK Biobank. BMD results for non-carriers ('NO', shown in blue) are compared to individuals with qualifying *COL1A1* variants ($N = 27$) and *COL1A2* variants ($N = 38$). Variant carriers are split into those without a clinical phenotype ('Yes', pink) and those with a clinical phenotype ('Yes – affected', red). For both clinically-affected groups, the BMD is significantly lower than across the UK Biobank as a whole. For *COL1A1*, BMD was also significantly reduced for variants carriers who did not have a clinical phenotype. Statistical comparisons between groups were performed using the Wilcoxon rank-sum test in RStudio. To aid visualisation, BMD results for non-carriers were randomly down-sampled before data points were plotted with the jitter function.

Two orthogonal datasets in UKB provided additional evidence of variant effects: circulating COL1A1 protein levels and BMD measurements. Whilst neither are clinically validated biomarkers, both provided additional utility. Decreased protein levels were confirmed in pLOF variant carriers (Fig. 3), with results also supporting the exclusion of a mosaic start-gain variant, present in 27.6% of reads (Fig. S3B). The modest correlation seen in participants for whom duplicate measurements were available, suggests inherent measurement noise and long-term environmental influences on COL1A2 levels (Note S2**)**. Although BMD measurements for lumbar spine are most relevant for assessing OI, this data was only available for <10% individuals. In contrast, heel BMD was available for >50% UKB participants and correlated with lumbar BMD. The group of individuals with qualifying *COL1A1* variants but no relevant ICD-10 code or evidence for multiple fractures had a significant decrease in heel BMD. This suggests that a variant may be 'non-penetrant' in terms of clinical phenotype but may still have a subclinical effect on bone density, highlighting that estimates of penetrance are highly dependent on the outcome measure being assessed. In contrast, BMD data supported low penetrance for *COL1A2* variants, as variant carriers with no relevant clinical findings had similar BMDs to the non-carrier group (Fig. 4). These results highlight the potential for companion 'omics datasets and biometric data to enhance penetrance/gNBS studies.

Somatic mosaicism of *COL1A1/2* variants can in rare cases help explain differences in expressivity [46–48]. Where mosaicism results in a milder skeletal phenotype, it is thought that variants must be present in >75% of osteoblasts [47]. In a population context where testing is limited to a single blood sample, it is impossible to assess relative tissue distributions. It was therefore prudent for variants that appear mosaic in blood to be excluded before assessing penetrance. In previous work, a VAF threshold of 27% balanced sensitivity and specificity for detecting somatic mosaicism in individuals with developmental disorders [49]. In penetrance studies, specificity is more important than sensitivity. Given that WGS in UKB is at a mean coverage of 32.5x, detection of somatic mosaicism would be limited. However, by combining WGS with the higher-coverage exome sequencing, we were able to remove several variants with VAF of 27-40% (Table 1). Examples of retained variants not quite passing the threshold for mosaicism are presented in Note S8. Combining mosaic with false-positive variant calls, a total of 24 *COL1A1* variants were excluded. Without manual review and mosaicism analysis, the increase in qualifying variants would have a major impact on the accuracy of downstream penetrance estimates. These results highlight the importance of undertaking a manual curation of variants [50]. Examples of multi-nucleotide variants, called as separate indel/SNVs (Fig. S3D); also highlight that care must be taken to avoid 'double-counting' individuals with complex variants.

The present study highlights several generalizable lessons that supplement our previous guidance on estimating disease penetrance [50]. For instance, four cases with qualifying variants in *COL1A1* (p.Gly1079Ser, p.Arg312Cys, p.Arg1014Cys and p.Pro535LeufsTer6) had multiple fractures reported and were initially considered to be likely affected individuals. However, closer scrutiny revealed that all fractures occurred within a short time frame and were concomitant with road traffic collisions or other physical accidents. This suggested the fractures may not be due to a genetic bone fragility condition and as no other evidence for bone fragility was identified, these individuals were classed as unaffected. Such cases illustrate how a systematic, chronological review of complete electronic clinical records can enhance the study of rare disease.

Although the UKB research analysis platform provides a wealth of clinical records to researchers, it is hard to estimate the effects of missing or incomplete clinical data. This is a major limitation of population cohort studies, compared with clinically ascertained rare-disease longitudinal cohort studies. However, given the relatively high penetrance for pLOF variants in *COL1A1* (>70%), this is unlikely to be a major factor. A recent study that examined 91 diseases linked to haploinsufficiency suggested that incomplete

penetrance for pLOF variants may be driven by residual allelic activity rather than by missing data [4]. Moreover, in the context of gNBS, where potential disease-causing variants would be reported within the first month of life, the presence of known P/LP variants in adults >40 years of age who were able to volunteer for a population study gives pause for thought. This finding implies that either the variants are incompletely penetrant and could result in substantial overdiagnosis if reported in the absence of a phenotype, or that individuals have milder or sub-clinical forms of disease that may be unnecessary to detect in newborns.

Like all large population cohorts, UK Biobank suffers from biases introduced by the volunteering process. In addition to recruiting individuals who are healthier than the average UK adult [34], this has introduced other sociodemographic biases [51]. For instance, the proportion of individuals self-reporting as white is increased, compared to data from the 2011 and 2021 UK censuses [52, 53]. Such biogeographical differences may influence genetic modifiers or environmental factors that modulate disease penetrance/expressivity. Extrapolation of our results to newborn screening programmes with participants from a different sociodemographic context should thus proceed with caution.

In conclusion, our study using WGS across a clinically unselected cohort of ~500,000 adults highlights the importance of using population data to inform gNBS, and the value of integrating proteomic/biometric data. Our results suggest the penetrance of OI in the general population is lower than for clinically ascertained cohorts, which is only explained in part by depletion of severe early-onset forms of the disease. For missense variants that alter the Gly-Xaa-Yaa repeat structure in α-collagen, the low penetrance suggests that over-reliance on current ClinVar assertions to support pathogenicity may also contribute to the overestimation of OI risk in gNBS programmes.

DATA AVAILABILITY

Data from UK Biobank cannot be shared publicly because of data availability and data return policies. Data are available from the UK Biobank for researchers who meet the criteria for access to datasets to UK Biobank (www.ukbiobank.ac.uk). Code used as part of this study is available at https://lcpilling.github.io/ukbrapR (R package for working in the UK Biobank Research Analysis Platform), https://github.com/drarwood/vep_ukb_aou_docker (VEP Annotation Process using Docker), or is to be described elsewhere. Further details are available on request.

REFERENCES

1. Wilson JMG, Jungner G, World Health O. Principles and practice of screening for disease. Geneva: World Health Organization; 1968.
2. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013;132:1077–130.
3. Mirshahi UL, Colclough K, Wright CF, Wood AR, Beaumont RN, Tyrrell J, et al. Reduced penetrance of MODY-associated HNF1A/HNF4A variants but not GCK variants in clinically unselected cohorts. Am J Hum Genet. 2022;109:2018–28.
4. Blair DR, Risch N. Reduced penetrance is common among predicted loss-of-function variants and is likely driven by residual allelic activity. medRxiv. 2025:2024.09.23.24314008.
5. Minten T, Bick S, Adelson S, Gehlenborg N, Amendola LM, Boemer F, et al. Data-driven consideration of genetic disorders for global genomic newborn screening programs. Genet Med. 2025;27:101443.
6. Genomic England Generation Study [Available from: www.generationstudy.co.uk.
7. Marini JC, Forlino A, Bachinger HP, Bishop NJ, Byers PH, Paepe A, et al. Osteogenesis imperfecta. Nat Rev Dis Prim. 2017;3:17052.
8. Rodriguez Celin M, Steiner RD, Basel D. COL1A1- and COL1A2-Related Osteogenesis Imperfecta. In: Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews((R)). Seattle (WA) 1993.
9. Etich J, Lessmeier L, Rehberg M, Sill H, Zaucke F, Netzer C, et al. Osteogenesis imperfecta-pathophysiology and therapeutic options. Mol Cell Pediatr. 2020;7:9.
10. Lindahl K, Astrom E, Rubin CJ, Grigelioniene G, Malmgren B, Ljunggren O, et al. Genetic epidemiology, prevalence, and genotype-phenotype correlations in the Swedish population with osteogenesis imperfecta. Eur J Hum Genet. 2015;23:1042–50.
11. England G. Genomics England PanelApp: A crowdsourcing tool to allow gene panels to be shared, downloaded, viewed and evaluated by the Scientific Community 2026 [Available from: https://panelapp.genomicsengland.co.uk/panels/196/.
12. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51:1560–5.
13. Australia P. Osteogenesis Imperfecta and Osteoporosis (Version 1.17). 2025.
14. England N. The national genomic test directory 2025 [version 8.1:[Available from: www.england.nhs.uk/publication/national-genomic-test-directories.
15. Maioli M, Gnoli M, Boarini M, Tremosini M, Zambrano A, Pedrini E, et al. Genotype-phenotype correlation study in 364 osteogenesis imperfecta Italian patients. Eur J Hum Genet. 2019;27:1090–100.
16. Cho TJ, Lee KE, Lee SK, Song SJ, Kim KJ, Jeon D, et al. A single recurrent mutation in the 5'-UTR of IFITM5 causes osteogenesis imperfecta type V. Am J Hum Genet. 2012;91:343–8.
17. Thornley P, Bishop N, Baker D, Brock J, Arundel P, Burren C, et al. Non-collagen pathogenic variants resulting in the osteogenesis imperfecta phenotype in children: a single-country observational cohort study. Arch Dis Child. 2022;107:486–90.
18. Bella J, Eaton M, Brodsky B, Berman HM. Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. Science. 1994;266:75–81.
19. Sillence DO, Senn A, Danks DM. Genetic heterogeneity in osteogenesis imperfecta. J Med Genet. 1979;16:101–16.
20. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.
21. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021;599:628–34.
22. Consortium UKBW-GS. Whole-genome sequencing of 490,640 UK Biobank participants. Nature. 2025.
23. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.
24. Cho SY, Lee JH, Ki CS, Chang MS, Jin DK, Han HS. Osteogenesis imperfecta Type I caused by a novel mutation in the start codon of the COL1A1 gene in a Korean family. Ann Clin Lab Sci. 2015;45:100–5.
25. Dalgleish R. The human type I collagen mutation database. Nucleic Acids Res. 1997;25:181–7.
26. Korkko J, Kuivaniemi H, Paassilta P, Zhuang J, Tromp G, DePaepe A, et al. Two new recurrent nucleotide mutations in the COL1A1 gene in four patients with osteogenesis imperfecta: about one-fifth are recurrent. Hum Mutat. 1997;9:148–56.
27. Information NCfB. ClinVar database 2026 [Available from: https://www.ncbi.nlm.nih.gov/clinvar/.
28. de Sainte Agathe JM, Filser M, Isidor B, Besnard T, Gueguen P, Perrin A, et al. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. Hum Genomics. 2023;17:7.
29. Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the Integrative Genomics Viewer. Cancer Res. 2017;77:e31–e4.
30. Trust SCsNF. Metabolic Bone Disease Service 2026 [Available from: www.sheffieldchildrens.nhs.uk/services/metabolic-bone-disease-service.
31. Organization WH. International Statistical Classification of Diseases and Related Health Problems 10th Revision [Available from: https://icd.who.int/browse10/2019/en.
32. Dhindsa RS, Burren OS, Sun BB, Prins BP, Matelska D, Wheeler E, et al. Rare variant associations with plasma protein levels in the UK Biobank. Nature. 2023;622:339–47.
33. Thompson DJ, Wells D, Selzam S, Peneva I, Moore R, Sharp K, et al. A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. PLoS One. 2024;19:e0307270.
34. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. Am J Epidemiol. 2017;186:1026–34.
35. Graves LE, Wall CL, Briody JN, Bennetts B, Wong K, Onikul E, et al. High bone mineral density osteogenesis imperfecta in a family with a novel pathogenic variant in COL1A2. Horm Res Paediatr. 2020;93:263–71.
36. Lindahl K, Barnes AM, Fratzl-Zelman N, Whyte MP, Hefferan TE, Makareeva E, et al. COL1 C-propeptide cleavage site mutations cause high bone mass osteogenesis imperfecta. Hum Mutat. 2011;32:598–609.
37. Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature. 2018;562:268–71.

38. Zheng SL, Jurgens SJ, McGurk KA, Xu X, Grace C, Theotokis PI, et al. Evaluation of polygenic scores for hypertrophic cardiomyopathy in the general population and across clinical settings. Nat Genet. 2025;57:563–71.
39. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405–24.
40. Kaplanis J, Deen D, Sivakumar P, de Almeida Gomes M, Puzriakova A, Leong I, et al. Assessment of the variant prioritisation strategy for genomic newborn screening in the Generation Study. Genet Med. 2025;27:101532.
41. Byers PH, Steiner RD. Osteogenesis imperfecta. Annu Rev Med. 1992;43:269–82.
42. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562:217–22.
43. Fowler DM, Rehm HL. Will variants of uncertain significance still exist in 2030? Am J Hum Genet. 2024;111:5–10.
44. Rivera-Munoz EA, Milko LV, Harrison SM, Azzariti DR, Kurtz CL, Lee K, et al. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. Hum Mutat. 2018;39:1614–22.
45. Investigators GPP, Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, et al. 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. N Engl J Med. 2021;385:1868–80.
46. Edwards MJ, Wenstrup RJ, Byers PH, Cohn DH. Recurrence of lethal osteogenesis imperfecta due to parental mosaicism for a mutation in the COL1A2 gene of type I collagen. The mosaic parent exhibits phenotypic features of a mild form of the disease. Hum Mutat. 1992;1:47–54.
47. Cabral WA, Marini JC. High proportion of mutant osteoblasts is compatible with normal skeletal function in mosaic carriers of osteogenesis imperfecta. Am J Hum Genet. 2004;74:752–60.
48. Cohen-Solal L, Zolezzi F, Pignatti PF, Mottes M. Intrafamilial variable expressivity of osteogenesis imperfecta due to mosaicism for a lethal G382R substitution in the COL1A1 gene. Mol Cell Probes. 1996;10:219–25.
49. Wright CF, Prigmore E, Rajan D, Handsaker J, McRae J, Kaplanis J, et al. Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. Nat Commun. 2019;10:2985.
50. Wright CF, Sharp LN, Jackson L, Murray A, Ware JS, MacArthur DG, et al. Guidance for estimating penetrance of monogenic disease-causing variants in population cohorts. Nat Genet. 2024;56:1772–9.
51. van Alten S, Domingue BW, Faul J, Galama T, Marees AT. Reweighting UK Biobank corrects for pervasive selection bias due to volunteering. Int J Epidemiol. 2024;53:dyae054.
52. Biobank U. UK Biobank data showcase: data-field 21000 "ethnic background" 2025 [updated August 2025. Available from: https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21000.
53. GOV.UK. Population of England and Wales 2022 [updated 21 May 2024. Available from: www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales/latest.

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource under Application Number 103356 and uses data provided by patients and collected by the NHS as part of their care and support. The authors would like to acknowledge the use of the University of Exeter High-Performance Computing (HPC) facility in carrying out this work.

AUTHOR CONTRIBUTIONS

ATP analysed data and wrote the manuscript. ELB, LJ and CFW conceived and supervised the study. RNB provided statistical advice and assisted with variant analysis. DB, SK and MB helped with variant curation, clinical phenotypes in OI and contributed to manuscript writing. JF and TH wrote custom scripts for data analysis.

FUNDING

This work was supported by the Medical Research Council [MR/X021351/1] and the National Institute for Health and Care Research Exeter Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

COMPETING INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI IN THE MANUSCRIPT PREPARATION PROCESS

During the preparation of this work, the authors used ChatGPT (OpenAI; GPT-5.2, knowledge cutoff August 2025) to assist with grammar correction and improve readability. All content generated or refined with this tool was reviewed, edited, and approved by the authors, who take full responsibility for the final manuscript.

ETHICAL APPROVAL

UK Biobank protocols were approved by the National Research Ethics Service Committee and all participants were recruited following informed consent. Data access was obtained under the approved UKB project 'Understanding the role of rare and common genetic variation in human phenotypes', application number 103356.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41431-026-02118-6.

Correspondence and requests for materials should be addressed to Caroline F. Wright.