

Do P-Wave Velocity-UCS Equations Transfer Across Studies? How Many Local Tests Are Needed for Intact-Rock Calibration

Ali Mehrabifard^{1,2}

Ali.Mehrabifard@ubc.ca

University of British Columbia

Zahra Yousefabad³

AIM Consulting Associates Inc.

Research Article

Keywords: P-wave velocity, uniaxial compressive strength, intact rock, external validation, local calibration, weak rock

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10856819/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

P-wave velocity (V_p) is frequently used to estimate intact-rock uniaxial compressive strength (UCS), but empirical equations calibrated in one study often fail when transferred to new rocks or laboratories. To quantify this transferability, 19 published relations were evaluated against 2,337 paired laboratory measurements from 59 distinct studies using leave-one-study-out validation. The findings demonstrate severe limitations in generalized equations: 14 out of 18 eligible equations failed to outperform a baseline benchmark that completely ignored velocity. Localized study-to-study shifts account for 62% of the predictive variance, and not a single published equation remained statistically valid for weak rocks below 25 MPa. While integrating four local UCS tests improved correct strength-class predictions from 45.0% to 63.2%, the remaining uncertainty was too large to select a precise design strength. Therefore, V_p should be restricted to preliminary core classification and screening, with direct local UCS measurements remaining mandatory wherever weak materials or numerical strength values dictate engineering design.

1 Introduction

Uniaxial compressive strength (UCS) is central to intact-rock characterization, but its measurement is destructive and requires suitable core. P-wave velocity (V_p) is faster to obtain, so empirical relations of the form $UCS = aV_p^b$ are often used to screen core, interpolate between tested intervals, or provide preliminary strength estimates. The difficulty is transfer: most relations are fitted to one study and then used on material that was not part of that calibration.

A high coefficient of determination within the source study does not establish that the same equation will predict a new study. Velocity and strength both respond to porosity, mineralogy, fabric, saturation, weathering, specimen geometry, and test procedure (Toksöz et al. 1979; Klimentos and McCann 1990; ISRM 2007). These factors can shift the velocity-strength relation without producing an obvious change in its overall slope. The relevant question is therefore not only whether velocity and strength correlate, but how much predictive information a published equation retains when applied to an independent study.

Previous compilations already point to this distinction. Chang et al. (2006) summarized relations for sedimentary rocks, while Aladejare (2020) and Aladejare et al. (2021) showed that index-based UCS models are commonly site dependent. Ching et al. (2018) evaluated transformation models against the ROCK/9/4069 database and found that performance on pooled data was generally poorer than source-study fits implied. In geotechnical reliability, this loss of transfer is treated as transformation uncertainty, distinct from inherent variability and measurement error (Phoon and Kulhawy 1999a, 1999b). Hierarchical models provide a natural way to represent it because they separate a common relation from variation among studies (Bozorgzadeh et al. 2019).

Transfer failure does not necessarily make velocity unhelpful. A common relation can be adjusted with a small local UCS dataset, as proposed in analogous site-specific and quasi-site-specific transformation frameworks (Wang and Aladejare 2015; Ching et al. 2021; Wu et al. 2022). The practical value of that adjustment depends on what it is expected to deliver. A few tests may be enough to improve strength classification while remaining inadequate for a precise numerical strength estimate. Likewise, an interval for one additional specimen and uncertainty in a locally calibrated mean level answer different questions and should not be used interchangeably.

Weak rock deserves separate attention. Material below about 25 MPa often governs excavation, slope, and underground design decisions, yet it occupies a small part of the published database. Apparent overprediction at

the weak end may reflect a genuine model deficiency, sparse sampling, or an artefact created by grouping residuals according to the observed strength being predicted. A useful assessment must distinguish among these possibilities and report what the available data can and cannot resolve.

This study assembles 2,337 paired laboratory measurements from 59 study-resolved datasets and asks four related questions. First, do published velocity-strength equations outperform matched benchmarks when evaluated outside their source studies and under the most specific source-supported restrictions available? Second, is the principal transfer error associated with functional form or with systematic differences among studies? Third, how much do small local UCS datasets improve prediction and strength classification? Fourth, what evidence is available for weak rock? Published equations are evaluated by leave-one-study-out validation and study-level resampling. A hierarchical model then separates the common velocity dependence, between-study offsets, and within-study scatter. Local calibration is tested only on held-out studies, and every claimed benefit is compared with an alternative that does not use velocity.

The scope is deliberately narrow. All paired measurements used in the analysis are from prepared intact specimens. The results do not validate borehole sonic velocities, rock-mass strength, or a design strength for a field site. They test whether laboratory V_p contributes useful information for intact-rock screening and local calibration, and how much direct UCS testing is still required before a numerical strength estimate can be defended.

2 Published velocity-strength relations

Twenty-four published relations predicting uniaxial strength from velocity alone were catalogued. Relations requiring additional index tests, using nonparametric or fuzzy learners rather than closed-form equations, or predicting properties other than strength were catalogued but not scored. They either require inputs beyond velocity or fall outside the velocity-to-strength benchmark (Mishra and Basu 2013; Siratovich et al. 2014; Heidari et al. 2018; Wang et al. 2020). Of the twenty-four, five are not scored: one requires Schmidt rebound as a second predictor, two are algebraic inversions of a regression fitted the other way round, of which one is Yaşar and Erdoğan (2004), one has a velocity unit that could not be recovered from the source, and one has an incomplete coefficient set. The nineteen scored relations come from seventeen source papers, two of which each contribute a pair.

The catalogue records, for each relation, its functional form, stated lithology, calibrated velocity range, and reported calibration fit. For two entries, fit statistics recorded in the source notes were external test results from later papers rather than calibration statistics. Because the original sources could not be obtained, those values were retained only as external-test statistics and the calibration-fit fields were left empty. The comparison between reported fit and transfer performance in Section 5.1 therefore includes sixteen equations. Acquisition method was not used as a moderator because it could not be verified consistently across all source papers.

Only two of the nineteen state a calibrated velocity range explicitly. Ten are assigned the observed velocity range of their own source study where that study appears in the corpus, and seven state neither and are left unrestricted in velocity. Where the restriction is inactive it is recorded as such per equation.

3 Database

3.1 Assembly and quality control

The analysed corpus contains 2,337 paired V_p and UCS measurements from 59 study-resolved datasets. It combines two evidence streams. The first comprises 385 specimen-level records from 13 studies extracted for this project. Searches were run in Scopus, Web of Science, and Google Scholar using four predefined combinations of velocity, UCS, intact rock, and laboratory-testing terms; backward and forward citation searches were then applied to eligible papers. Of 55 records identified, five duplicates were removed, 50 titles and abstracts were screened, 28 full texts were assessed, and 13 studies provided recoverable paired specimen data. The second stream comprises 1,952 records from 46 study identifiers inherited from ROCK/10/4025 (Muzamhindo and Ferentinou 2023), a CC BY 4.0 compilation. These inherited records were not independently selected through the project search. The work is therefore described as a benchmark assembled from a targeted search and an existing database rather than as a systematic review of the full literature.

Assembly yielded 4,108 candidate pairs. Four quality-control screens removed 1,133 records, and a further 638 records without study-resolved source identifiers were excluded. The remaining 2,337 records formed the analysed corpus (Table 1). Each candidate record, its final inclusion status, and the reason for every exclusion are reported in Supplementary Data 1. Five of the 46 inherited identifiers, representing 246 analysed records, are generic source labels rather than author-year citations. They remain distinct study units in the source map; Section 5.6 reports a sensitivity analysis excluding them.

The four exclusion screens removed duplicate specimens, substantially overlapping study units, values outside prespecified bounds of 500 to 8,500 m/s and 0.5 to 700 MPa, and manufactured materials. Records without a study-resolved source identifier were also excluded because location labels could not be treated reliably as study units. Every exclusion and its reason are reported in Supplementary Data 1.

Two non-exclusion checks flagged internal inconsistencies and suspiciously repeated strength values. Seven studies, covering 585 records, exceeded their study-specific collision null at $p < 0.01$. Because rounding can produce or conceal such collisions, none was removed; Section 5.6 reports the refit without all seven. Four repeated author names were also confirmed as separate papers from disjoint measurements. Merging those pairs would increase the offset standard deviation from 0.526 to 0.537.

Table 1
Assembly and quality control of the corpus. Every candidate record is listed with its final status and, when excluded, its reason in Supplementary Data 1.

Stage	Records
Paired velocity and strength records assembled	4,108
Duplicate measurement of a specimen already present	-1,065
Study unit substantially contained in another	-35
Outside prespecified plausibility bounds	-18
Manufactured material rather than intact rock	-15
Retained after quality control	2,975
Without a study-resolved source identifier	-638
Analysed	2,337 in 59 studies

3.2 What the corpus contains

Strength runs from 0.68 to 560.3 MPa with a median of 70.0 MPa, and velocity from 509 to 7,943 m/s with a median of 4,580 m/s (Fig. 1). Sedimentary rocks supply 1,186 records from 44 studies, igneous 771 from 26, and metamorphic 379 from 21; one record is unclassified (Fig. 1; Supplementary Data 1). Study counts sum to more than 59 because a study reporting more than one lithology contributes to each class. Studies are unequal in size, holding from 3 to 185 records with a median of 29, and the ten largest supply 47.2 percent of the corpus, which is why every metric below weights studies rather than records. They are also narrow: the median study covers a factor of 6.0 in strength where the corpus covers a factor of 824, so the pooled relation is fitted across a range no single laboratory programme observed.

Rock type is reported by the source papers in 126 distinct raw strings, including case variants, weathering-grade concatenations and one evident typographical error. Only the three-way genetic class is harmonised; the raw strings are carried unaltered and reported in Supplementary Data 1.

On the ISRM field-strength descriptors (ISRM 2007), the corpus is concentrated in strong rock. Class R4, 50 to 100 MPa, holds 819 records from 48 studies and class R5, 100 to 250 MPa, holds 693 from 43, together 64.7 percent of the corpus. The weak end is thin: R2, 5 to 25 MPa, holds 292 records from 34 studies, R1 holds 80 from 5 studies, and R0 holds 7 from a single study. R0 and R1 together are 3.7 percent of the corpus.

The record-level data underlying these summaries are provided in Supplementary Data 1.

3.3 The corpus is sparse at low strength and at velocity extremes

Sorted into 1,000 m/s velocity bands, the corpus is concentrated in the middle (Fig. 2; Supplementary Data 1). The three bands between 3,500 and 6,500 m/s hold 1,538 records, 65.8 percent of the total. At or below 1,500 m/s there are 104 records from 15 studies, and above 6,500 m/s there are 120 records from 11 studies. Bands are closed on the right throughout the paper, so a record at exactly a band edge belongs to the lower band.

Within each band the strength range is wide, and the width is the point. In the 2,500 to 3,500 m/s band the interquartile range of strength runs from 19.7 to 62.9 MPa against a full range of 1.5 to 145 MPa, so a velocity measured to that band is consistent with rock spanning three ISRM classes. Only in the slowest band does velocity constrain strength tightly, and that band is carried by 15 studies.

Several mechanisms may contribute to this imbalance. Weak, porous and clay-rich rock attenuates the compressional wave more strongly than dense competent rock (Toksöz et al. 1979; Klimentos and McCann 1990), which can make first arrivals harder to pick. Specimen preparation and reporting may also filter weak material, although this corpus cannot separate or quantify those effects. Whatever the cause, velocity data are sparse in the material whose strength may govern design, and Section 5.5 examines the resulting evidence gap.

4 Methods

4.1 The pooled relation

Fitting used $\log V_p$ and $\log UCS$. This linearizes the power relation and produces approximately symmetric residual scatter. The pooled least-squares relation over all 2,337 records is

$$\text{UCS} = a V_p^b \quad (1)$$

with $a = 2.762 \times 10^{-5}$ and $b = 1.7505$, giving a coefficient of determination of 0.627 in log units.

No smearing correction is applied when predictions are returned to MPa. A Duan (1983) smearing factor estimates the conditional arithmetic mean, whereas the point predictions and intervals in this study are defined on the conditional median. Applying the factor would put the point estimate and its interval on different targets. As a sensitivity check, applying the empirical factor of 1.1879 changed the study-balanced held-out error by +1.36 MPa (paired 95 percent cluster-bootstrap interval -1.07 to $+3.86$ MPa; $p = 0.27$), providing no evidence of improved transfer.

4.2 The hierarchical model

A pooled fit treats every record as exchangeable, which these data are not. The model used throughout gives each study its own offset from the common relation:

$$\log \text{UCS}_{ij} = \beta_0 + \beta_1 \log V_{p,ij} + u_j + \epsilon_{ij} \quad (2)$$

$$u_j \sim \mathcal{N}(0, \tau^2), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (3)$$

for record i in study j . Here, β_0 is the intercept, β_1 is the common log-velocity slope, u_j is the study offset, and ϵ_{ij} is the within-study residual. The two variance components separate systematic offsets among studies from residual variation among specimens within a study. In the vocabulary of Phoon and Kulhawy (1999a), τ represents the transformation uncertainty encountered when the pooled relation is transferred to a new study population, while σ is the residual specimen-level variation. These components cannot isolate laboratory procedure from lithology, sampling, saturation, or other unreported differences among studies. Bozorgzadeh et al. (2019) fitted a hierarchical Bayesian regression of this shape to 40 granite strength datasets from the literature, and what is done here applies that method to the velocity-strength pair. Estimation is by Hamiltonian Monte Carlo with a non-centred parameterisation of the offsets. The prior on τ is a half-normal with scale 0.5, and six alternatives were run as a sensitivity.

Predicting a study absent from the training set requires one fresh offset draw per posterior sample, shared across every record of that study. Drawing an offset per record instead would convert between-study variance into independent noise, leaving marginal coverage looking correct while destroying the within-study correlation.

4.3 Validation

Leave-one-study-out is the only headline validation scheme. Each of the 59 studies is withheld in turn, the model is refitted on the remaining 58, and the withheld study is predicted as an unseen study population. Random record-level cross-validation is not reported as a headline because records within a study are not independent, and a scheme that splits them inflates apparent skill by letting the model see the offset it is supposed to predict (Roberts et al. 2017).

Metrics are study-balanced, so that a 185-record study cannot decide the result for a corpus whose typical study holds a few dozen. For RMSE, error is computed within each held-out study and then averaged with equal weight across studies. Interval performance is scored with the interval score of Gneiting and Raftery (2007), which penalises width, so that a model cannot win by being vague.

Coverage is assessed against its nominal level, while calibration error is the absolute deviation from nominal after study-balanced averaging. Every inferential comparison uses a 4,000-replicate cluster bootstrap that resamples studies, with paired resampling for matched conditions. Skill-ratio intervals use bias-corrected and accelerated limits because the ratio distribution is skewed at small numbers of clusters (Efron 1987). A study drawn repeatedly is removed from its own leave-one-study-out training pool for each occurrence. Bootstrap p values are two-sided and limited by the resampling resolution. Holm correction is applied within comparison families; prespecified single contrasts are reported without multiplicity adjustment.

4.4 The calibration experiment

Each study is held out in turn, m of its specimens are drawn as the local calibration set, the pooled relation is corrected by their mean residual, and the correction is scored on the study's remaining specimens. The calibration specimens are never scored, the target study never contributes to the global relation or to the variance components, the evaluation set is held fixed as m grows, and the spread across random draws is reported because a practitioner gets one draw rather than an average.

Four velocity-based selection rules are compared: random unstructured selection, selection spanning the local velocity range, selection of the slowest specimens first, and velocity-spanning selection with the slowest specimen forced into the set. At $m = 1$ the final two coincide. Comparisons across m use the same 23-study cohort for every rule. The weak-rock analysis in Section 5.5 instead uses all 37 eligible studies, only six of which are weak-rock dominated. That small subset cannot resolve differences among selection rules, so cohort sizes and intervals are reported throughout.

4.5 Two accuracy criteria taken from practice

Two quantities are reported because they answer different practical questions. The first is the uncertainty in the locally calibrated study offset under the common-slope model. Ruffolo and Shakoor (2009) considered relative-error targets of 10, 15, 20, and 25 percent at 95 percent confidence when sizing laboratory programmes. Using the same target levels for comparison, the posterior standard deviation of the local offset after m UCS tests is

$$s_m = \sqrt{\frac{\tau^2 \sigma^2}{m\tau^2 + \sigma^2}} \quad (4)$$

The corresponding upper relative margin is $\exp(1.96s_m) - 1$. It describes uncertainty in the conditional strength level associated with the local offset, not uncertainty in a physical site-wide mean or in a design strength. The interval for one further specimen is wider, with standard deviation $\sqrt{\sigma^2 + s_m^2}$, and has an irreducible floor at σ . Both calculations use the posterior medians of τ and σ .

The test counts required to reach each target are compared with two no-velocity baselines. One uses local UCS tests and the corpus grand mean. The other also conditions on the raw rock-type label reported within each study. This second grouping is a proxy for information that may be available before testing, but it is not a validation of a geological map or a field-scale geological model. Counts from these model-based comparisons indicate the relative information supplied by velocity and reported rock type; they are not universal sample-size prescriptions for a project.

The second quantity is resolution of an ordered strength class. The predictive density for each held-out study is integrated over the ISRM boundaries (ISRM 2007). Classes representing less than five percent of the corpus are

merged with an adjacent class, producing four categories with boundaries at 25, 50, and 100 MPa. The merged lowest category does not distinguish R0, R1, and R2, and the majority-class reference rate is 34.6 percent on the 37-study classification cohort.

The ranked probability score is used as the primary descriptive metric because it evaluates the full ordered predictive distribution and gives partial credit to near misses (Epstein 1969; Gneiting and Raftery 2007). Correct-class rate is reported for interpretation, and quadratic-weighted kappa is retained for comparison with earlier literature. This ordering was adopted after initial analyses showed that changes in predicted class marginals made kappa insensitive to the calibration contrast; the classification analysis is therefore explicitly exploratory. All three metrics are reported across m . Because kappa is undefined when every held-out record occupies one class, each metric is compared on the intersection of studies for which it is defined, with the cohort size reported.

4.6 Scoring the published equations

Each equation is scored both unrestricted and under the most specific recoverable restrictions: positive predictions, matching declared genetic class, and a stated or source-study velocity range. Only two equations report an explicit velocity range; ten use their source-study range and seven remain unrestricted in velocity. Known or probable source-study matches are removed before scoring; the equation-level match list is available from the corresponding author on reasonable request. A common-set sensitivity scores all equations after excluding the union of those matches.

For each equation and bootstrap replicate, the mean-only benchmark is recomputed from the other studies in the applicable test set because the restrictions produce different test sets. A second benchmark conditions on harmonised genetic class and falls back to the grand mean when necessary. It represents coarse lithological information available without velocity, not a field-scale geological model. Section 5.1 reports both comparisons.

5 Results

5.1 Published equations carry little transferable information

Unrestricted, the best relation, Karakul and Ulusay (2013), has a study-balanced root-mean-square error (RMSE) of 39.6 MPa, compared with 53.4 MPa for the mean-only benchmark. The median equation has an RMSE of 49.1 MPa, and 11 of 19 equations have lower point-estimate error than the benchmark (Table 2; Fig. 3).

Applying the available equation-specific restrictions reduces the median RMSE from 48.7 to 39.1 MPa, but it also reduces the matched benchmark from 53.4 to 45.0 MPa. Median skill consequently changes from $+0.087$ to $+0.030$, with 11 equations exceeding the benchmark in both analyses (Wilcoxon $p = 0.91$, $k = 18$).

The uncertainty intervals provide the clearest summary. Unrestricted, seven of 19 equations have 95 percent skill intervals wholly above zero and six wholly below it. Under the available restrictions, four of 18 intervals are wholly above zero and none is wholly below zero. These four are Tuğrul and Zarif (1999), Karakul and Ulusay (2013), Moradian and Behnia (2009), and Kahraman (2001), with point skills of $+0.30$, $+0.26$, $+0.21$, and $+0.20$, respectively. Holm correction retains all four: adjusted $p < 0.001$ for the latter three and $p = 0.045$ for Tuğrul and Zarif. The remaining 14 equations are not distinguishable from the matched benchmark.

The restrictions retain very different portions of the corpus. Karakul and Ulusay retain all 2,337 records and 59 studies; Kahraman retains 2,098 records in 57 studies; Moradian and Behnia retain 1,159 records in 44 studies; and

Tuğrul and Zarif retain 752 records in 26 studies. The last two are the materially restricted cases. Across the full catalogue, restriction removes the poorest extrapolations, but it does not turn uncertain equations into transferable ones.

A second benchmark conditioned on harmonised genetic class changes the unrestricted benchmark only from 53.37 to 52.49 MPa and reduces the number of equations beating it under the available restrictions on the point estimate from 11 to 10. The four supported relations have RMSE values of 35.7 to 40.7 MPa, compared with matched no-velocity benchmarks of 48.7 to 53.4 MPa. Against a corpus median UCS of 70.0 MPa, this is useful but still imprecise transfer.

The largest observed errors in these examples are associated with extrapolation. Diamantis et al. (2011) reaches 408.5 MPa unrestricted and has zero test records under its available restrictions, so that number describes a relation calibrated on serpentinites and peridotites near 5,000 m/s being evaluated far outside that band. Khandelwal and Singh (2009) moves from 281.6 MPa to 25.8 MPa on the seventeen studies retained under its restrictions. Negative predicted strengths, 5,506 of them across the unrestricted catalogue, fall to zero once each linear relation is held above its own zero crossing. The finding that survives restriction is that most of these equations still do not beat a mean-only benchmark.

A pooled power law refitted on all 2,337 records has a study-balanced RMSE of 39.70 MPa in leave-one-study-out prediction, which the best published equation essentially ties. This result suggests that the main transfer limitation is not the precision with which the pooled curve is fitted. Section 5.5 tests richer functional forms directly.

Because each relation is scored on a different subset once restriction is applied, the ranking could be an artefact of which studies each is graded on. Rescoring all nineteen on the common set of 50 studies and 2,107 records that every relation can be evaluated on preserves it: eleven of nineteen beat the benchmark unrestricted and eleven of eighteen under the available restrictions, with median skills of $+0.085$ and $+0.037$ against benchmarks of 55.9 and 44.2 MPa.

Calibration R^2 does not identify the equations that transfer. Across the 16 relations with a reported fit, its rank correlation with study-balanced external error is $+0.43$ ($p = 0.10$; 95 percent bootstrap interval -0.14 to $+0.81$).

Table 2

Published equations scored outside their calibration studies, unrestricted and under the most specific equation-specific restrictions available. Restrictions combine declared lithology with a stated or source-study velocity band where recoverable; an unavailable restriction is inactive. Study-balanced RMSE gives each study equal weight. Skill is one minus the ratio of equation error to the matched mean-only error; zero is the benchmark and one is exact prediction. Intervals are 95 percent bias-corrected and accelerated cluster-bootstrap intervals. The benchmark is recomputed for every cell. Restriction retains is the percentage of the scoring corpus remaining after the available restrictions are applied. Calibration R^2 is the source-paper value; n.r. denotes not reported.

Equation	Calib. R^2	Unrestricted RMSE (MPa)	Unrestricted skill [95% CI]	Restriction retains	Restricted studies	Restricted RMSE (MPa)	Restricted skill [95% CI]
Karakul and Ulusay (2013)	0.78	39.6	+ 0.26 [+ 0.16, + 0.35]	100%	59	39.6	+ 0.26 [+ 0.16, + 0.35]
Kahraman (2001)	0.69	41.5	+ 0.22 [+ 0.12, + 0.31]	91%	57	40.7	+ 0.20 [+ 0.11, + 0.29]
Ghazvinian et al. (2005)	n.r.	42.7	+ 0.20 [+ 0.08, + 0.30]	50%	44	43.4	+ 0.11 [-0.09, + 0.24]
Azimian et al. (2014)	0.91	42.8	+ 0.20 [+ 0.08, + 0.30]	50%	44	43.6	+ 0.11 [-0.10, + 0.24]
Moradian and Behnia (2009)	0.70	43.1	+ 0.19 [+ 0.10, + 0.28]	50%	44	38.5	+ 0.21 [+ 0.10, + 0.31]
Abbaszadeh Shahri et al. (2016), updated Kahraman	0.85	45.1	+ 0.15 [+ 0.01, + 0.27]	19%	29	26.0	-0.09 [-0.49, + 0.18]
Mohamad et al. (2015)	0.83	45.3	+ 0.15 [+ 0.01, + 0.26]	15%	33	19.7	+ 0.06 [-0.12, + 0.18]
Tuğrul and Zarif (1999)	0.80	47.5	+ 0.11 [-0.06, + 0.24]	32%	26	35.7	+ 0.30 [+ 0.10, + 0.44]
McNally (1990)	n.r.	48.3	+ 0.09 [-0.08, + 0.23]	47%	43	51.6	-0.06 [-0.36, + 0.14]
Hosseini et al. (2013)	n.r.	49.1	+ 0.08 [-0.10, + 0.21]	47%	43	52.4	-0.08 [-0.36, + 0.13]
Khandelwal (2013)	0.87	50.8	+ 0.05 [-0.14, + 0.19]	3%	17	22.8	+ 0.00 [-1.12, + 0.12]
Abbaszadeh Shahri et al. (2016), direct	0.83	53.8	-0.01 [-0.24, + 0.17]	19%	29	26.2	-0.10 [-0.49, + 0.17]

Equation	Calib. R^2	Unrestricted RMSE (MPa)	Unrestricted skill [95% CI]	Restriction retains	Restricted studies	Restricted RMSE (MPa)	Restricted skill [95% CI]
Jamshidi et al. (2018)	0.82	57.4	-0.06 [-0.24, + 0.05]	13%	34	41.5	+ 0.03 [-0.06, + 0.10]
Çobanoğlu and Çelik (2008)	0.67	68.9	-0.28 [-0.48, -0.11]	13%	39	38.6	+ 0.03 [-0.07, + 0.08]
Kurtuluş et al. (2012), along foliation	0.83	76.3	-0.42 [-0.58, -0.28]	5%	16	68.9	-0.53 [-1.07, + 0.10]
Kurtuluş et al. (2012), across foliation	0.93	82.2	-0.53 [-0.79, -0.29]	5%	16	41.8	+ 0.07 [-0.22, + 0.45]
Sharma and Singh (2008)	0.90	97.2	-0.82 [-1.26, -0.38]	4%	15	24.8	-0.18 [-1.00, + 0.15]
Khandelwal and Singh (2009)	0.96	281.6	-4.27 [-5.50, -2.67]	3%	17	25.8	-0.13 [-1.06, + 0.14]
Diamantis et al. (2011)	0.83	408.5	-6.60 [-7.91, -4.55]	0%	0	no retained records	n.a.

5.2 Between-study offsets dominate transfer uncertainty

Two related statistics describe the heterogeneity, but they answer different questions. The meta-analytic statistic is $I^2 = 99.47$ percent ($Q = 10$, 850, 58 degrees of freedom). In the hierarchical predictive model, the posterior median standard deviations are $\tau = 0.5205$ log units for study offsets and $\sigma = 0.4073$ for within-study residuals, closely matching restricted maximum-likelihood estimates of 0.519 and 0.407. These variance components imply that 62 percent of observation-level predictive variance for a further record from an unobserved study is attributable to the study offset. This 62 percent is not an alternative estimate of I^2 .

The fitted offsets generate 59 near-parallel relations (Fig. 4). Their strength multipliers range from 0.21 to 3.97 relative to the pooled relation. The standard-deviation multiplier is 1.68 among studies, compared with 1.50 for residual variation within a study.

The reported study characteristics explain little of the offset (Fig. 5). Genetic class removed 10.28 percent of offset variance, compared with a 3.45 percent analytic baseline. Its 10.20 percent threshold was the 95th percentile from 10,000 permutations. Velocity range removed 6.59 percent (baseline 1.92 percent; threshold 7.43 percent), while density basis removed 4.09 percent (baseline 2.50 percent; threshold 9.38 percent). Genetic class is the only one of 11 moderators to exceed that threshold, and only by 0.08 percentage points. Its raw permutation p value is 0.049, but the Holm-adjusted value is 0.54 and the Benjamini-Hochberg value is $q = 0.37$.

Combining all available moderators explains about 14 percent of offset variance and reduces the plug-in 95 percent predictive span from a factor of 13.29 to 11.9. The posterior-median model gives a span of 13.34 without local information; the plug-in and posterior-median calculations agree within 1.4 percent across the tested calibration sizes.

Moderator analyses of per-study slope and intercept do not change this result. Velocity range is the closest slope moderator ($q = 0.051$), consistent with poorer slope identification in studies covering a narrow velocity range; no intercept moderator survives false-discovery correction (minimum $q = 0.21$).

Goodness-of-fit tests gave mixed evidence for the Gaussian offset model. The Shapiro-Wilk test did not reject normality ($W = 0.968$, $p = 0.121$), whereas the Anderson-Darling statistic was marginally above its 5 percent critical value ($A^2 = 0.749$ versus 0.742). A Student- t offset model estimates $\nu = 7.1$ with a wide 95 percent credible interval of 1.8 to 40.0 and leaves σ essentially unchanged (0.4072 versus 0.4073). Its predictive spans differ from the Gaussian model by no more than 1.3 percent over the tested calibration sizes, so the simpler Gaussian form is retained.

This result is bounded by what the literature reports, and that bound is itself a finding. The corpus carries no method moderators at all: no saturation state, no length-to-diameter ratio, no transducer frequency, no standard followed, no weathering grade. Those fields exist for the records extracted specimen by specimen for this project, and even there they are too sparse to regress on, with saturation state present for 42 of 385 records and porosity for 21. The offset is not shown to be unconditionable in principle. It is shown to be unconditionable on what published papers print, which is the situation a practitioner is in.

The complete moderator table, including permutation-null percentiles and multiplicity-adjusted values, is available from the corresponding author on reasonable request.

5.3 Information gained from local testing

Local testing is evaluated against three distinct quantities: the predictive interval for one further record, uncertainty in the locally calibrated conditional strength level, and exploratory strength-class resolution (Table 3).

The 95 percent predictive interval for one further record spans a factor of 13.34 with no local tests, 7.63 after one test, 5.82 after four, and 5.13 after 20 (Fig. 6). Its lower limit is about a factor of 4.94 because within-study residual variation remains after the study offset is known. The 95 percent relative margin for the locally calibrated conditional level is 177 percent with no tests, 87 percent after one, 45 percent after four, 28 percent after ten, and 19 percent after 20. Under this model, reaching margins of 25, 20, 15, and 10 percent requires 13, 19, 33, and 70 local tests, respectively.

The required count changes with the information supplied before testing. With neither velocity nor class conditioning, the corresponding counts are 24, 35, 60, and 128. With velocity but no class conditioning they are 13, 19, 33, and 70; with reported-rock-type conditioning but no velocity they are 13, 19, 32, and 68; and with both they are 8, 12, 21, and 44. These are model-based comparisons within this corpus. They quantify the relative information in velocity and reported rock type, not the test programme required for a specific field site.

Class resolution improves quickly and then flattens. On the 37-study cohort, the ranked probability score falls from 0.410 with no tests to 0.332 after one, 0.288 after four, and 0.274 after ten. At four tests, the paired contrast with no testing is -0.122 (95 percent cluster-bootstrap interval -0.195 to -0.056 ; $p < 0.001$). Correct-class rate

increases from 45.0 percent to 58.2, 63.2, and 64.2 percent over the same sequence; the four-test gain is 18.1 percentage points (95 percent interval $+7.6$ to $+29.0$; $p < 0.001$). The probability assigned to the true class rises from 0.464 to 0.608 after four tests.

Ten tests improve the ranked probability score by a further 0.0146 relative to four tests (95 percent interval $+0.011$ to $+0.019$), while the correct-class rate increases by only 1.1 percentage points (interval -0.5 to $+2.5$). Quadratic-weighted kappa does not resolve the main contrast; its paired intervals include zero at every test count. The $m = 15$ and $m = 20$ rows use smaller cohorts and should not be compared directly with the 37-study rows.

The three quantities place four tests in context. They capture most of the early improvement in the exploratory classification curve, but leave a wide interval for an additional specimen and a 45 percent relative margin on the locally calibrated conditional level. A four-test set is therefore a pragmatic initial tranche for class screening and for directing further work, not a universal optimum or a design-strength programme. A marginal-gain rule also selects four tests at its nominal 2 percent threshold, but selects between two and nine as that heuristic threshold varies from 6 to 0.5 percent; it is treated as a description of the curve rather than independent confirmation.

Table 3

Effect of m local UCS tests on three distinct quantities. Before merging, R0, R1, and R6 individually represented less than 5 percent of the corpus; R0 and R1 were merged upward with R2, and R6 downward with R5, giving four classes with boundaries at 25, 50, and 100 MPa. The conditional-level margin and the interval for one further specimen are not interchangeable: the latter retains within-study residual variation. Classification intervals are 95 percent cluster-bootstrap intervals over studies (4,000 resamples). Cohort sizes are shown because kappa is undefined for a study containing only one held-out class and because the 15- and 20-test rows include fewer eligible studies.

m	Conditional level, 95% rel. margin	One specimen, span (factor)	Class RPS [95% CI]	Correct class [95% CI]	Weighted kappa [95% CI]	Studies, RPS	Studies, kappa
0	177%	13.34	0.410 [0.357, 0.465]	45.0% [37.5, 52.8]	0.375 [0.291, 0.459]	37	36
1	87%	7.63	0.332 [0.283, 0.382]	58.2% [52.5, 64.5]	0.398 [0.323, 0.474]	37	36
2	64%	6.54	0.311 [0.262, 0.360]	60.8% [54.8, 67.2]	0.431 [0.349, 0.509]	37	35
3	52%	6.07	0.299 [0.248, 0.348]	62.0% [55.5, 68.7]	0.424 [0.341, 0.505]	37	35
4	45%	5.82	0.288 [0.239, 0.336]	63.2% [56.6, 69.9]	0.425 [0.342, 0.509]	37	35
5	40%	5.66	0.286 [0.237, 0.334]	63.8% [57.3, 70.4]	0.431 [0.344, 0.518]	37	35
10	28%	5.31	0.274 [0.225, 0.322]	64.2% [57.0, 71.5]	0.415 [0.313, 0.513]	37	34
15	22%	5.19	0.273 [0.222, 0.326]	63.2% [55.1, 71.3]	0.452 [0.342, 0.561]	29	27
20	19%	5.13	0.284 [0.226, 0.341]	60.8% [52.3, 69.7]	0.519 [0.421, 0.621]	23	21
50	12%	5.01	n.a.	n.a.	n.a.	n.a.	n.a.
100	8%	4.98	n.a.	n.a.	n.a.	n.a.	n.a.

5.4 Specimen selection matters as well as sample count

Selection rules were compared on the same fixed cohort of 23 studies, with paired study-level resampling and Holm correction across the eight tested calibration sizes (Table 4; Table 5; Fig. 7). At four tests, velocity-spanning selection produced 0.850 realised coverage of a nominal 0.80 interval, compared with 0.812 for unstructured selection. The increase reflected wider intervals, not better calibration. Spanning reduced median absolute offset

error by 0.099 log units at four tests (raw $p = 0.025$), but this contrast did not survive Holm correction ($p = 0.20$). The data do not establish an advantage over unstructured selection.

The consistent adverse result is concentration at the slow end. Slowest-first selection under-covers at every calibration size: realised coverage is 0.661 after one test, 0.697 after four, and 0.795 after 20, compared with 0.804 to 0.824 under unstructured selection. At four tests the paired coverage deficit is 0.115 (95 percent interval -0.190 to -0.048 ; Holm-adjusted $p = 0.002$). The deficit remains significant through ten tests (Holm-adjusted $p = 0.042$ at $m = 10$). This comparison supports avoiding a calibration set drawn entirely from low-velocity material; it does not show that the slowest material should be excluded from the broader strength-testing programme.

Reserving one low-velocity specimen while selecting the remainder across the velocity range avoids the all-slowest rule, but it is indistinguishable from unstructured selection at four tests on span, interval score, calibration error, and median absolute offset error. The comparisons therefore do not identify a universally optimal selection rule. They support one precaution: do not estimate the study offset from a set concentrated entirely at the slow end. Four tests remain a pragmatic exploratory classification tranche because most of the early class-resolution gain occurs by that point, not because the selection experiment identifies four as a universal optimum.

Table 4

Four-test held-out calibration results on the fixed cohort of 23 studies. Lower is better for span, calibration error, interval score, and median absolute offset error. Full results for every calibration size and all 95 percent intervals are available from the corresponding author on reasonable request.

Selection rule	Span (factor)	80% coverage	Calibration error	Interval score	Median abs. offset error
Unstructured	5.00	0.812	0.012	2.432	0.156
Velocity-spanning	4.52	0.850	0.050	2.304	0.057
Slowest-first	7.13	0.697	0.103	3.047	0.218
Spanning plus slowest	5.39	0.778	0.022	2.254	0.180

Table 5

Key paired contrasts at four local tests on the fixed 23-study cohort. Intervals are 95 percent cluster-bootstrap intervals from 4,000 paired resamples. Holm correction is across the eight calibration sizes tested for the same contrast; the full contrast grid is available from the corresponding author on reasonable request.

Contrast	Difference [95% CI]	Raw p	Holm p
Velocity-spanning minus unstructured, calibration error	+ 0.038 [-0.034, + 0.059]	0.311	1.000
Velocity-spanning minus unstructured, median abs. offset error	-0.099 [-0.134, -0.008]	0.025	0.200
Slowest-first minus unstructured, 80% coverage	-0.115 [-0.190, -0.048]	< 0.001	0.002

5.5 Weak rock: what the corpus can and cannot establish

The corpus contains 375 records from 35 studies below 25 MPa and 1,005 records from 52 studies below 60 MPa. The 25 MPa threshold is the upper boundary of ISRM class R2 (ISRM 2007); 60 MPa is retained only as a practice-relevant reference from weak-ground observations (Potvin and Hadjigeorgiou 2008), not as a universal definition. Evidence below 5 MPa is sparse: 86 records from five studies.

The pooled relation is not demonstrably biased in weak rock. Outcome-binned ratios appear to show overprediction, including a predicted-to-observed median ratio of 1.79 from 5 to 25 MPa. Binning on observed strength, however, also bins on the residual being assessed. An exactly unbiased relation simulated with the fitted between-study and within-study variance produces the same pattern. The observed 5-to-25 MPa ratio lies within its artefact-only range of 1.30 to 1.99, and no observed band exceeds its own upper null limit (Table 6; Fig. 8).

When a two-fold bias is injected into every specimen below 3,500 m/s, the test detects it in the 5-to-25 MPa band 95 percent of the time and in the 1-to-5 MPa band 69 percent of the time. When the same bias is confined below 2,500 m/s, power falls to 11 and 58 percent, respectively. It is also low for effects below about 1.5-fold. Analyses binned by velocity or predicted strength likewise do not establish systematic overprediction. The supported conclusion is therefore limited: the corpus does not demonstrate a large universal bias below 25 MPa, but it cannot exclude moderate or more localized bias, especially at the lowest strengths.

More flexible forms do not improve transfer. A quadratic term increases held-out weak-rock error by 0.0674 log units ($p < 0.001$), and a segmented model increases it by 0.0829 ($p < 0.001$). Neither reduces weak-band bias. The additional curvature improves in-sample flexibility but does not predict new studies.

Published equations provide no supported weak-rock prediction. Thirteen retain observations below 25 MPa under the available restrictions; none outperforms its matched mean-only benchmark. Their median absolute error is 29.1 MPa, compared with a weak-band benchmark error of 7.24 MPa. This result, rather than the outcome-binned bias pattern, is the strongest evidence against transferring a published equation into weak material without local testing.

The apparent decline in residual scatter from weak to strong rock is also reproduced when a homoscedastic relation is binned on observed strength. Binning on velocity removes the trend. One velocity band, 1,500 to 2,500 m/s, has greater scatter than its null expectation, but the slowest band does not. The corpus therefore does not establish a general increase in scatter with decreasing strength.

Only six of the 37 calibration studies are weak-rock dominated, so their intervals are too wide to rank selection rules. Local calibration narrows their median spans, but no rule can be preferred. The evidence gap is therefore substantive: five studies below 5 MPa and 15 studies below 1,500 m/s cannot support a general weak-rock calibration strategy. Where low-strength material governs a decision, its strength should be measured directly.

Table 6

Corpus composition and pooled-relation behaviour by ISRM strength class. ISRM class rows are closed on their upper boundary, whereas the combined rows use strict “below” thresholds; one record at exactly 5 MPa and four at exactly 25 MPa therefore remain in the lower ISRM class but are excluded from the corresponding strict-threshold count. The ratio is median predicted divided by median observed UCS. The artefact-only interval is the 95 percent range obtained when outcomes are regenerated from an exactly unbiased relation with the fitted between-study and within-study variance. Because classes are defined by observed UCS, ratios and residual coefficients of variation in this table are descriptive and cannot by themselves establish a weak-rock effect.

Class (MPa)	Records	Studies	Sed / Ig / Met	Median actual	Median predicted	Ratio	Artefact-only null	Within-study resid. COV
R0, $UCS \leq 1$	7	1	7 / 0 / 0	0.8	2.0	2.40	[1.71, 5.46]	17.0%
R1, $1 < UCS \leq 5$	80	5	74 / 0 / 6	2.4	5.3	2.19	[0.78, 2.56]	28.7%
R2, $5 < UCS \leq 25$	292	34	216 / 51 / 25	14.7	26.3	1.79	[1.30, 1.99]	34.6%
R3, $25 < UCS \leq 50$	437	44	238 / 133 / 66	38.1	50.0	1.31	[1.17, 1.70]	32.7%
R4, $50 < UCS \leq 100$	819	48	392 / 286 / 140	72.9	72.2	0.99	[0.98, 1.21]	28.2%
R5, $100 < UCS \leq 250$	693	43	253 / 298 / 142	134.6	98.2	0.73	[0.60, 0.73]	20.3%
R6, $UCS > 250$	9	2	6 / 3 / 0	395.8	97.2	0.25	[0.28, 0.40]	17.4%
UCS < 25, combined	375	35	295 / 49 / 31	12.2				
UCS < 60, combined	1005	52	629 / 256 / 120	32.1				

5.6 Robustness

Across 59 leave-one-study-out folds, realised coverage of the nominal 95 percent interval averages 0.950. Calibration is not uniform: 41 folds have complete coverage, while three cover less than half of their held-out observations (0.459, 0.308, and 0.286). Those three studies are also furthest from the pooled relation. The average therefore describes aggregate performance, not a guarantee for every new study. Fold-level R^2 is not summarized because many held-out studies span too little strength for it to be stable.

The primary full-data fit initially had a maximum $\hat{R}$ of 1.0109. A longer confirmation fit reduced it to 1.0019, increased the minimum bulk effective sample size from 303 to 1,272, and changed τ by 0.00011 and σ by 0.000002. No divergent transitions occurred in the leave-one-study-out sweep. Fold-level convergence was less uniform: 20 of 59 fits had a reported maximum $\hat{R}$ above 1.01, generally 1.02 to 1.03. The foldwise posterior summaries are therefore treated as approximate diagnostics rather than precise study-specific estimates. The main variance components and calibration curves are supported by the converged full fit and the method-of-moments sensitivity. The classification and specimen-selection experiments use separate REML fits within each held-out fold and do not depend on these MCMC fold summaries. Heavy-tailed offset fits also had low energy diagnostics and are used only to test the direction and magnitude of sensitivity, not as sources of headline estimates.

Fitting the project-extracted and inherited records separately gives exponents of 1.7280 and 1.7477, compared with 1.7505 combined. Their offset and residual standard deviations differ by less than 0.05 log units. When the 19 equations are rescored only on the 385 project-extracted records, their rank ordering correlates 0.933 with the full-corpus ordering ($p = 5.5 \times 10^{-9}$). The smaller subset cannot support the moderator analysis because most fields are too sparse.

Excluding the five generic inherited study identifiers moves the exponent from 1.7505 to 1.7769 and the held-out error from 41.43 to 41.48 MPa on a common test set. Excluding the seven studies flagged for repeated values also leaves the central conclusions unchanged.

Removing the six studies whose deletion most improves the pooled fit raises log-space R^2 from 0.627 to 0.702 but does not improve held-out error on the same 53 test studies (40.17 versus 40.03 MPa). The apparent improvement therefore reflects selective exclusion rather than better transfer.

6 Discussion

Study-level offsets, rather than an inadequately fitted common curve, are the main obstacle to transfer. The meta-analytic heterogeneity statistic is $I^2 = 99.47$ percent, while the hierarchical predictive decomposition attributes 62 percent of observation-level variance to study offsets. These statistics answer different questions: the former compares heterogeneity with sampling error, whereas the latter partitions predictive variance for a further observation.

Local UCS tests can estimate the offset, but they do not make velocity a precise substitute for strength testing. In the exploratory four-class analysis, four tests reduced the ranked probability score by 30 percent and increased correct classification from 45.0 to 63.2 percent. The 95 percent span for one further specimen narrowed from a factor of 13.34 to 5.82, but a 45 percent relative margin remained on the locally calibrated conditional level. Four tests are therefore a pragmatic starting point for class screening and targeting subsequent work, not a universal optimum or a sufficient programme for selecting a design strength.

Model-based test counts reinforce this distinction. Velocity and reported rock-type conditioning each reduce uncertainty in this corpus, and using both reduces it further. These comparisons measure the information carried by the available variables; they do not validate mapped field units or prescribe a sample size for every project. A local programme should be sized from the decision and its required uncertainty, with velocity used to stratify and guide testing rather than to replace it.

The interval for one further specimen has an irreducible floor set by within-study scatter. Local calibration reduces uncertainty in the study offset but cannot remove specimen-level variation. This is why classification can improve appreciably while numerical predictions remain broad.

Weak rock remains the principal evidence gap. Outcome-binned plots suggest overprediction of weak-rock strength, but the same pattern appears under an unbiased null and therefore does not by itself establish bias in the pooled velocity-strength relation. Power for a two-fold bias in the 5-to-25 MPa band is high only when the injected effect extends below 3,500 m/s; it falls sharply when the effect is confined below 2,500 m/s and remains limited for moderate or localized bias. More flexible pooled curves do not improve held-out weak-rock prediction.

The clearest comparative result concerns weak rock: none of the 13 published equations retaining records below 25 MPa under the available restrictions outperforms the matched mean-only benchmark. Together with the small number of studies below 5 MPa, this supports a conservative practical boundary. Where weak material affects design, its strength should be measured locally. Future datasets should give low-strength material greater representation and report the population to which each relation is intended to apply.

The offset may represent irreducible transfer uncertainty or unreported moderators. The results do not show that saturation, transducer frequency, specimen geometry, weathering, or test procedure are unimportant; they show that the literature does not report them consistently enough to explain the offset. The contribution is therefore a quantified local-calibration workflow in the spirit of site-specific transformation models (Bozorgzadeh et al. 2019; Ching et al. 2021), together with an explicit account of what the available calibration data can and cannot deliver.

Limitations. Most records were inherited from an existing compilation rather than independently extracted from primary papers, and five inherited study identifiers lack author-year metadata. Seven studies have a repeated-value signature. Sensitivity analyses show little effect on the central estimates, but they cannot replace source-level verification. The calibration experiment treats a published study as a target population; a field site or production core run may be more heterogeneous. The 23-study selection cohort is biased toward larger datasets, only six eligible studies are weak- rock dominated, and fold-level Bayesian convergence is incomplete. Seven equations declare no velocity range, making their restricted evaluations comparatively permissive.

The conclusions apply to paired measurements on prepared intact specimens. They do not establish a laboratory-to-borehole velocity conversion, rock-mass strength, or a design value for a field site. Applying the workflow to sonic logs would require co-registered core, laboratory velocity, UCS, and log data so that the velocity translation and the local strength offset can both be validated.

7 Conclusions

Published P-wave velocity–UCS equations are not generally transferable estimates of intact strength. Of the 18 equations retaining data under the available restrictions, only four outperform the matched no-velocity benchmark

after family-wise correction; the other 14 show no detectable improvement over it. A study-level offset dominates transfer error and is not explained by the reported study characteristics after multiplicity correction.

A small local dataset can still be useful. In the exploratory four-class analysis of 37 held-out studies, four UCS tests reduce the ranked probability score by 30 percent, increase correct classification from 45.0 to 63.2 percent, and narrow the 95 percent interval for one further specimen from a factor of 13.34 to 5.82. Even after calibration, however, the conditional strength level retains a 45 percent relative margin. Four tests are therefore a practical starting point for screening and targeting further tests, not a universal optimum or a substitute for a design-strength programme.

Selection analysis supports one clear precaution: do not calibrate solely on the lowest-velocity material. Beyond that, the data do not reliably rank unstructured, velocity-spanning, and mixed strategies. The local programme should represent the material relevant to the engineering decision and expand until the required uncertainty is reached.

Weak-rock inference is especially limited. Below 25 MPa, no published equation retaining observations under the available restrictions outperforms the corresponding mean-only benchmark. The data are also too sparse to exclude moderate pooled-relation bias or to identify a preferred weak-rock calibration rule. Where low-strength material governs design, UCS should therefore be measured locally.

The transferable contribution is a workflow rather than a universal equation: use velocity to stratify and conditionally classify intact core within the local population, estimate the study-level offset with direct tests, and size the remaining UCS programme according to the uncertainty the engineering decision can tolerate. Future velocity-strength studies will be more transferable if they report saturation, specimen geometry, transducer frequency, test standard, and the velocity and strength ranges over which the relation is intended to apply.

Declarations

- **Supplementary Information.** Supplementary Data 1 contains all 4,108 candidate records, identifies the source paper or source-record label for each record, and states whether the record was included in or excluded from the analysis, with the reason for each exclusion.

Author Contributions. Ali Mehrabifard led the study, developed the concept, designed the methodology and analysis, conducted the software development, data curation, and formal analysis, prepared the figures and original manuscript, and supervised the work. Zahra Yousefabadi contributed to the methodology, validation, data curation, analysis, interpretation, visualization, and review and editing of the manuscript. Both authors read and approved the final manuscript.

Funding. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability. The corpus and complete screening record are provided as Supplementary Data 1. It contains all 4,108 candidate records, the source paper or source-record label, the final inclusion status, and the reason for every exclusion. Part of the corpus derives from ROCK/10/4025 (Muzamhindo and Ferentinou 2023) under CC BY 4.0 and is redistributed under the same licence with attribution.

Code Availability. All analysis code can be provided on reasonable request from the author.

Competing Interests. The authors have no relevant financial or non-financial interests to disclose.

Declaration of generative AI in the writing process. During preparation of this work, generative AI tools assisted with writing and language editing. The authors reviewed and revised the resulting text and take responsibility for the submitted content.

ORCID. Ali Mehrabifard: <https://orcid.org/0000-0003-3160-8812>.

List of symbols

List of symbols are available in the Supplementary Files section.

References

1. Abbaszadeh Shahri A, Larsson S, Johansson F (2016) Updated relations for the uniaxial compressive strength of marlstones based on P-wave velocity and point load index test. *Innov Infrastruct Solut* 1:17. <https://doi.org/10.1007/s41062-016-0016-9>
2. Aladejare AE (2020) Evaluation of empirical estimation of uniaxial compressive strength of rock using measurements from index and physical tests. *J Rock Mech Geotech Eng* 12(2):256–268. <https://doi.org/10.1016/j.jrmge.2019.08.001>
3. Aladejare AE, Alofe ED, Onifade M, Lawal AI, Ozoji TM, Zhang ZX (2021) Empirical estimation of uniaxial compressive strength of rock: database of simple, multiple, and artificial intelligence-based regressions. *Geotech Geol Eng* 39(6):4427–4455. <https://doi.org/10.1007/s10706-021-01772-5>
4. Azimian A, Ajalloeian R, Fatehi L (2014) An empirical correlation of uniaxial compressive strength with P-wave velocity and point load strength index on marly rocks using statistical method. *Geotech Geol Eng* 32:205–214. <https://doi.org/10.1007/s10706-013-9703-x>
5. Bozorgzadeh N, Harrison JP, Escobar MD (2019) Hierarchical Bayesian modelling of geotechnical data: application to rock strength. *Geotechnique* 69(12):1056–1070. <https://doi.org/10.1680/jgeot.17.p.282>
6. Chang C, Zoback MD, Khaksar A (2006) Empirical relations between rock strength and physical properties in sedimentary rocks. *J Pet Sci Eng* 51(3–4):223–237. <https://doi.org/10.1016/j.petrol.2006.01.003>
7. Ching J, Li KH, Phoon KK, Weng MC (2018) Generic transformation models for some intact rock properties. *Can Geotech J* 55(12):1702–1741. <https://doi.org/10.1139/cgj-2017-0537>
8. Ching J, Phoon KK, Ho YH, Weng MC (2021) Quasi-site-specific prediction for deformation modulus of rock mass. *Can Geotech J* 58(7):936–951. <https://doi.org/10.1139/cgj-2020-0168>
9. Çobanoğlu I, Çelik SB (2008) Estimation of uniaxial compressive strength from point load strength, Schmidt hardness and P-wave velocity. *Bull Eng Geol Environ* 67:491–498. <https://doi.org/10.1007/s10064-008-0158-x>
10. Diamantis K, Bellas S, Migiros G, Gartzos E (2011) Correlating wave velocities with physical and mechanical properties and petrographic characteristics of peridotites from the central Greece. *Geotech Geol Eng* 29(6):1049–1062. <https://doi.org/10.1007/s10706-011-9436-7>
11. Duan N (1983) Smearing estimate: a nonparametric retransformation method. *J Am Stat Assoc* 78(383):605–610. <https://doi.org/10.1080/01621459.1983.10478017>

12. Efron B (1987) Better bootstrap confidence intervals. *J Am Stat Assoc* 82(397):171–185.
<https://doi.org/10.1080/01621459.1987.10478410>
13. Epstein ES (1969) A scoring system for probability forecasts of ranked categories. *J Appl Meteorol* 8(6):985–987. [https://doi.org/10.1175/1520-0450\(1969\)008%3C0985:ASSFPF%3E2.0.CO;2](https://doi.org/10.1175/1520-0450(1969)008%3C0985:ASSFPF%3E2.0.CO;2)
14. Ghazvinian A, Fathi A, Rashidi M, Gharacheh M (2005) Application of nondestructive tests in measurements of some physical and mechanical parameters of Zagros marls. In: *Proceedings of the Mining Engineering Conference, Tehran, Iran*
15. Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 102(477):359–378. <https://doi.org/10.1198/016214506000001437>
16. Heidari M, Mohseni H, Jalali SH (2018) Prediction of uniaxial compressive strength of some sedimentary rocks by fuzzy and regression models. *Geotech Geol Eng* 36:401–412. <https://doi.org/10.1007/s10706-017-0334-5>
17. Hosseini M, Alipanahi B, Senemari S (2013) Determination of engineering properties of marlstones using punch test. *Appl Geol* 8(4):309–322
18. ISRM (2007) The complete ISRM suggested methods for rock characterization, testing and monitoring: 1974–2006. In: Ulusay R, Hudson JA (eds) *Suggested methods prepared by the Commission on Testing Methods, International Society for Rock Mechanics*. ISRM Turkish National Group, Ankara, 628 pp
19. Jamshidi A, Zamanian H, Zarei Sahamieh R (2018) The effect of density and porosity on the correlation between uniaxial compressive strength and P-wave velocity. *Rock Mech Rock Eng* 51(4):1279–1286. <https://doi.org/10.1007/s00603-017-1379-8>
20. Kahraman S (2001) Evaluation of simple methods for assessing the uniaxial compressive strength of rock. *Int J Rock Mech Min Sci* 38(7):981–994. [https://doi.org/10.1016/S1365-1609\(01\)00039-9](https://doi.org/10.1016/S1365-1609(01)00039-9)
21. Karakul H, Ulusay R (2013) Empirical correlations for predicting strength properties of rocks from P-wave velocity under different degrees of saturation. *Rock Mech Rock Eng* 46:981–999. <https://doi.org/10.1007/s00603-012-0353-8>
22. Khandelwal M (2013) Correlating P-wave velocity with the physico-mechanical properties of different rocks. *Pure Appl Geophys* 170(4):507–514. <https://doi.org/10.1007/s00024-012-0556-7>
23. Khandelwal M, Singh TN (2009) Correlating static properties of coal measures rocks with P-wave velocity. *Int J Coal Geol* 79(1–2):55–60. <https://doi.org/10.1016/j.coal.2009.01.004>
24. Klimentos T, McCann C (1990) Relationships among compressional wave attenuation, porosity, clay content, and permeability in sandstones. *Geophysics* 55(8):998–1014. <https://doi.org/10.1190/1.1442928>
25. Kurtuluş C, Bozkurt A, Endes H (2012) Physical and mechanical properties of serpentinized ultrabasic rocks in NW Turkey. *Pure Appl Geophys* 169(7):1205–1215. <https://doi.org/10.1007/s00024-011-0394-z>
26. McNally GH (1990) The prediction of geotechnical rock properties from sonic and neutron logs. *Explor Geophys* 21(1–2):65–71. <https://doi.org/10.1071/EG990065>
27. Mishra DA, Basu A (2013) Estimation of uniaxial compressive strength of rock materials by index tests using regression analysis and fuzzy inference system. *Eng Geol* 160:54–68. <https://doi.org/10.1016/j.enggeo.2013.04.004>
28. Mohamad ET, Jahed Armaghani D, Momeni E, Alavi Nezhad Khalil Abad SV (2015) Prediction of the unconfined compressive strength of soft rocks: a PSO-based ANN approach. *Bull Eng Geol Environ* 74(3):745–757. <https://doi.org/10.1007/s10064-014-0638-0>

29. Moradian ZA, Behnia M (2009) Predicting the uniaxial compressive strength and static Young's modulus of intact sedimentary rocks using the ultrasonic test. *Int J Geomech* 9(1):14–19. [https://doi.org/10.1061/\(ASCE\)1532-3641\(2009\)9:1\(14\)](https://doi.org/10.1061/(ASCE)1532-3641(2009)9:1(14))
30. Muzamhindo H, Ferentinou M (2023) Generic compressive strength prediction model applicable to multiple lithologies based on a broad global database. *Probabilistic Eng Mech* 71:103400. <https://doi.org/10.1016/j.pro bengmech.2022.103400>
31. Phoon KK, Kulhawy FH (1999a) Characterization of geotechnical variability. *Can Geotech J* 36(4):612–624. <https://doi.org/10.1139/t99-038>
32. Phoon KK, Kulhawy FH (1999b) Evaluation of geotechnical property variability. *Can Geotech J* 36(4):625–639. <https://doi.org/10.1139/t99-039>
33. Potvin Y, Hadjigeorgiou J (2008) Ground support strategies to control large deformations in mining excavations. *J South Afr Inst Min Metall* 108(7):397–404. <https://www.saimm.co.za/Journal/v108n07p397.pdf>
34. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Aroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40(8):913–929. <https://doi.org/10.1111/ecog.02881>
35. Ruffolo RM, Shakoor A (2009) Variability of unconfined compressive strength in relation to number of test samples. *Eng Geol* 108(1–2):16–23. <https://doi.org/10.1016/j.enggeo.2009.05.011>
36. Sharma PK, Singh TN (2008) A correlation between P-wave velocity, impact strength index, slake durability index and uniaxial compressive strength. *Bull Eng Geol Environ* 67:17–22. <https://doi.org/10.1007/s10064-007-0109-y>
37. Siratovich PA, Heap MJ, Villeneuve MC, Cole JW, Reuschle T (2014) Physical property relationships of the Rotokawa Andesite, a significant geothermal reservoir rock. *Geotherm Energy* 2:10. <https://doi.org/10.1186/s40517-014-0010-4>
38. Toksöz MN, Johnston DH, Timur A (1979) Attenuation of seismic waves in dry and saturated rocks: I. Laboratory measurements. *Geophysics* 44(4):681–690. <https://doi.org/10.1190/1.1440969>
39. Tuğrul A, Zarif IH (1999) Correlation of mineralogical and textural characteristics with engineering properties of selected granitic rocks from Turkey. *Eng Geol* 51(4):303–317. [https://doi.org/10.1016/S0013-7952\(98\)00071-4](https://doi.org/10.1016/S0013-7952(98)00071-4)
40. Wang M, Wan W, Zhao Y (2020) Prediction of the uniaxial compressive strength of rocks from simple index tests using a random forest predictive model. *C R Mecanique* 348(1):3–32. <https://doi.org/10.5802/crmeca.3>
41. Wang Y, Aladejare AE (2015) Selection of site-specific regression model for characterization of uniaxial compressive strength of rock. *Int J Rock Mech Min Sci* 75:73–81. <https://doi.org/10.1016/j.ijrmms.2015.01.008>
42. Wu S, Ching J, Phoon KK (2022) Quasi-site-specific soil property prediction using a cluster-based hierarchical Bayesian model. *Struct Saf* 99:102253. <https://doi.org/10.1016/j.strusafe.2022.102253>
43. Yaşar E, Erdoğan Y (2004) Correlating sound velocity with the density, compressive strength and Young's modulus of carbonate rocks. *Int J Rock Mech Min Sci* 41(5):871–875. <https://doi.org/10.1016/j.ijrmms.2004.01.012>

Figures

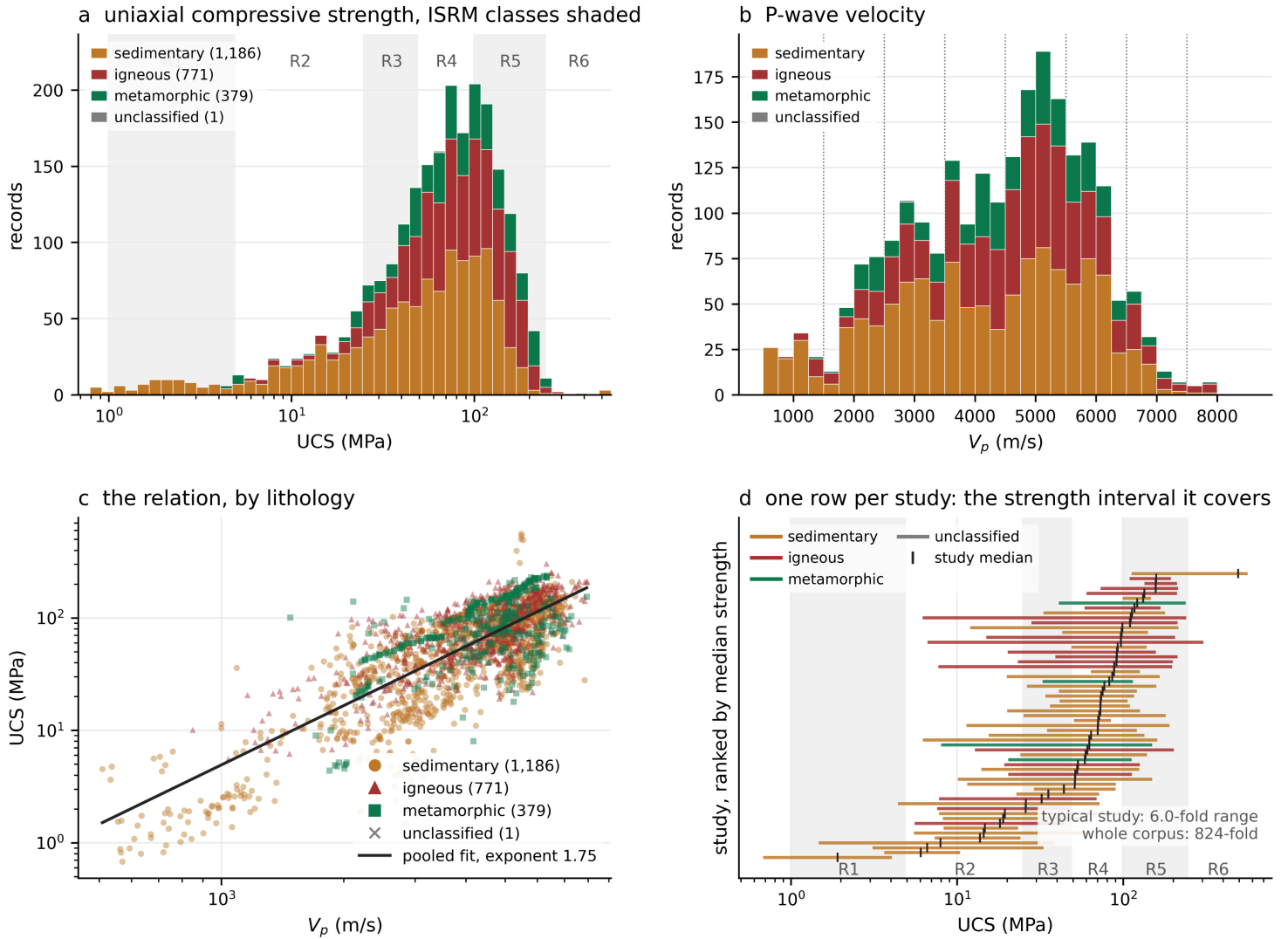


Figure 1

Anatomy of the analysed corpus. All four panels share one colour for genetic class. (a) Distribution of uniaxial compressive strength, stacked by genetic class, with the ISRM field-strength classes shaded; 295 of the 375 records below 25 MPa are sedimentary. (b) Distribution of P-wave velocity on the same stacking, with the band edges of Figure 2 marked. (c) The velocity-strength plane by genetic class with the pooled power relation. (d) The strength interval each of the 59 studies covers, sorted by study median, with the median marked; the median study spans a factor of 6.0 in strength where the corpus spans a factor of 824.

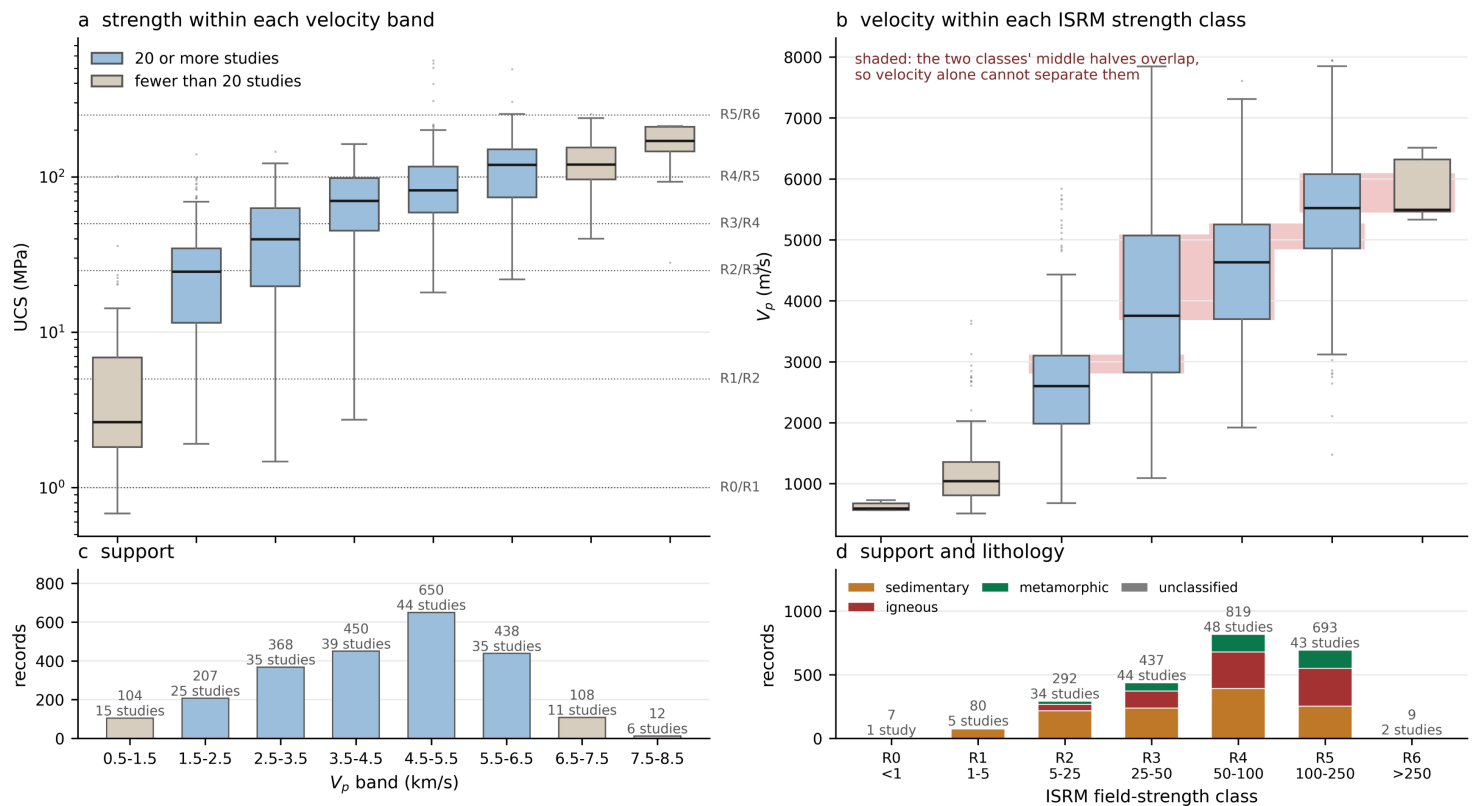


Figure 2

The same corpus read in both directions. (a) Strength within each 1,000 m/s velocity band and (c) the records and studies behind each band. (b) Velocity within each ISRM field-strength class and (d) the records, studies and lithology behind each class; shading in (b) marks adjacent classes whose interquartile velocity ranges overlap. Boxes carried by fewer than twenty studies are drawn in a lighter fill. Velocity materially narrows the strength range only in the slowest band, while the field-strength classes retain broad, overlapping velocity distributions.

4 of 18 restricted comparisons favour the published equation over a mean-only benchmark with 95% bootstrap intervals

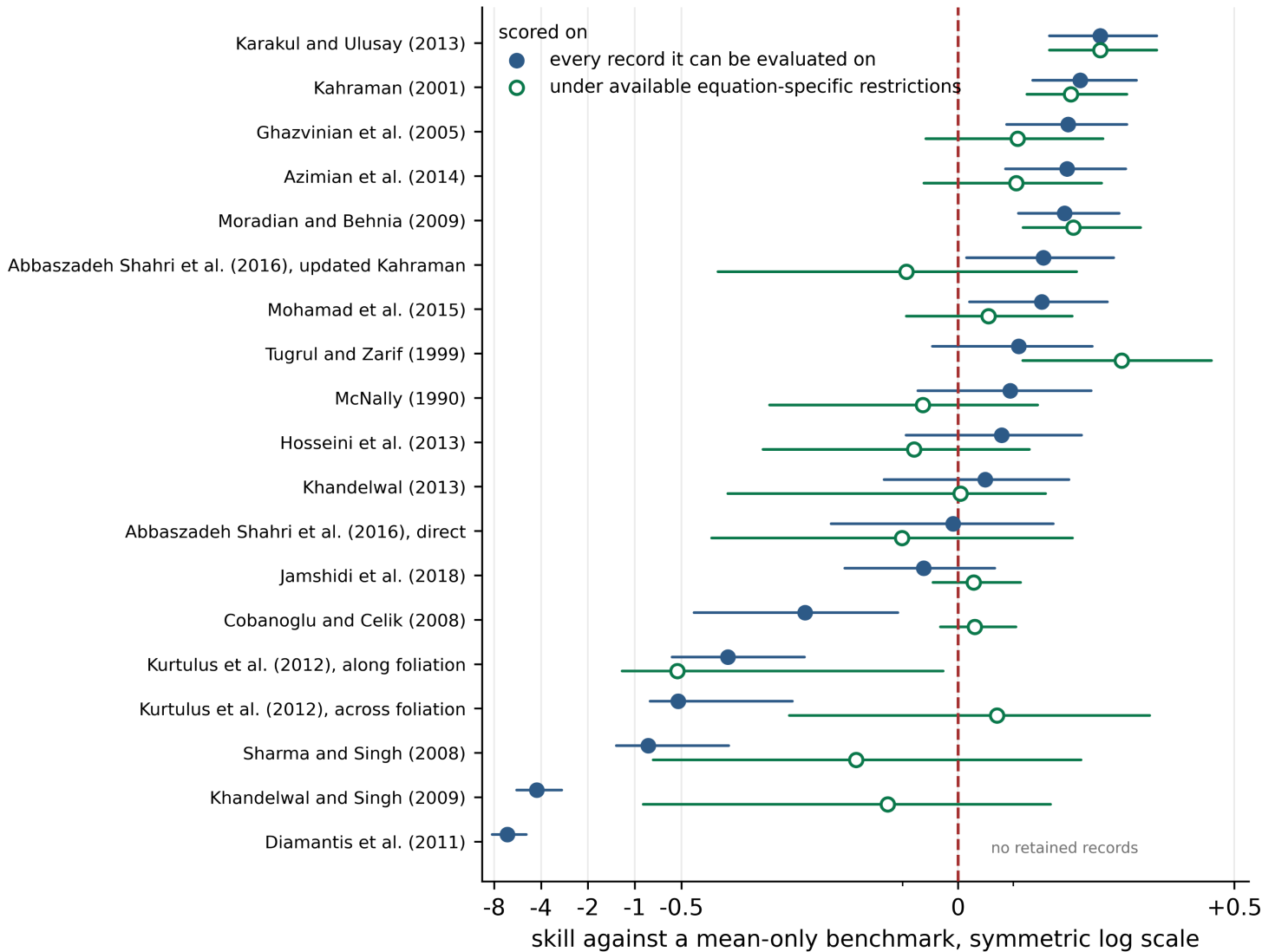


Figure 3

Skill of the nineteen published equations relative to a mean-only benchmark, scored with the records of their own calibration study withheld. Filled markers show unrestricted scoring; open markers apply the most specific equation-specific lithology and velocity restrictions available, with unavailable restrictions left inactive. Bars are 95 percent cluster-bootstrap intervals over studies. Four of 18 restricted intervals lie wholly right of zero. Diamantis et al. (2011) retains no test records under its restrictions and is shown unrestricted only.

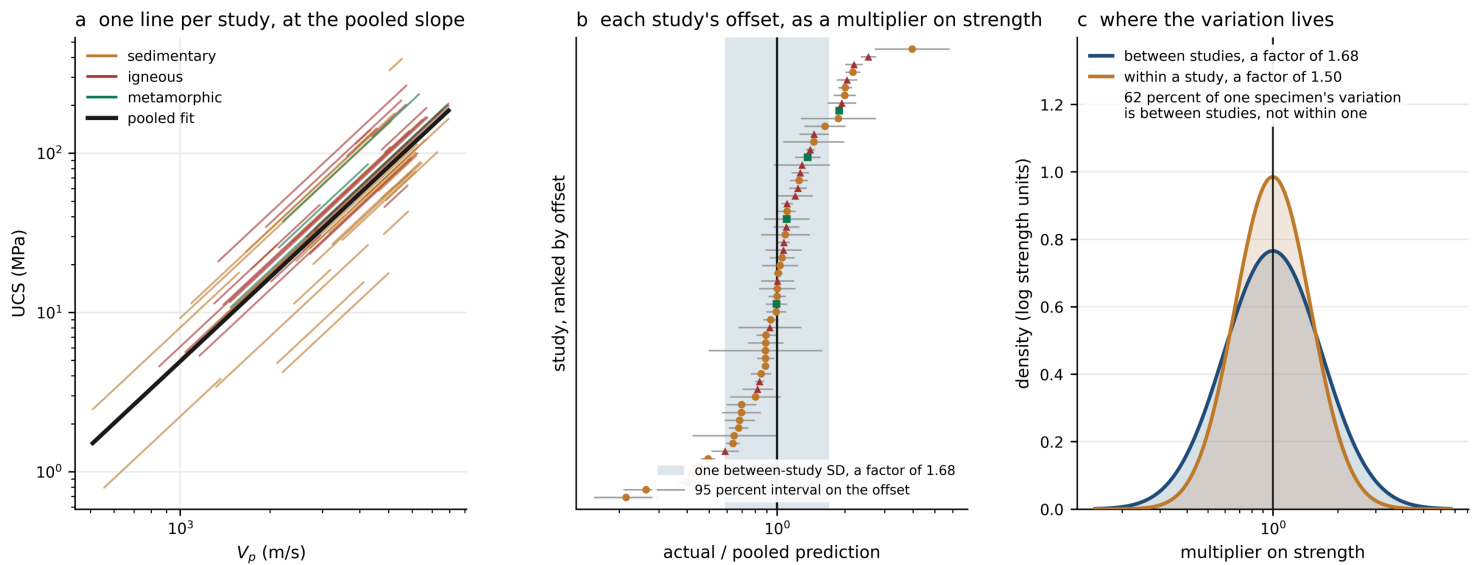


Figure 4

The study-level offset shown three ways. (a) The pooled slope displaced by each study's fitted offset and drawn over that study's measured velocity range. (b) Offsets expressed as multipliers on strength, with intervals based on the standard error of each study mean and one between-study standard deviation shaded. (c) Fitted between-study and within-study distributions on the same multiplicative scale. The offset multipliers range from 0.21 to 3.97, and study offsets account for 62 percent of observation-level predictive variance for a further record from an unobserved study.

The strongest reported characteristic, rock class, clears its own null at $p = 0.049$, and at $q = 0.37$ once the 11 tried are accounted for

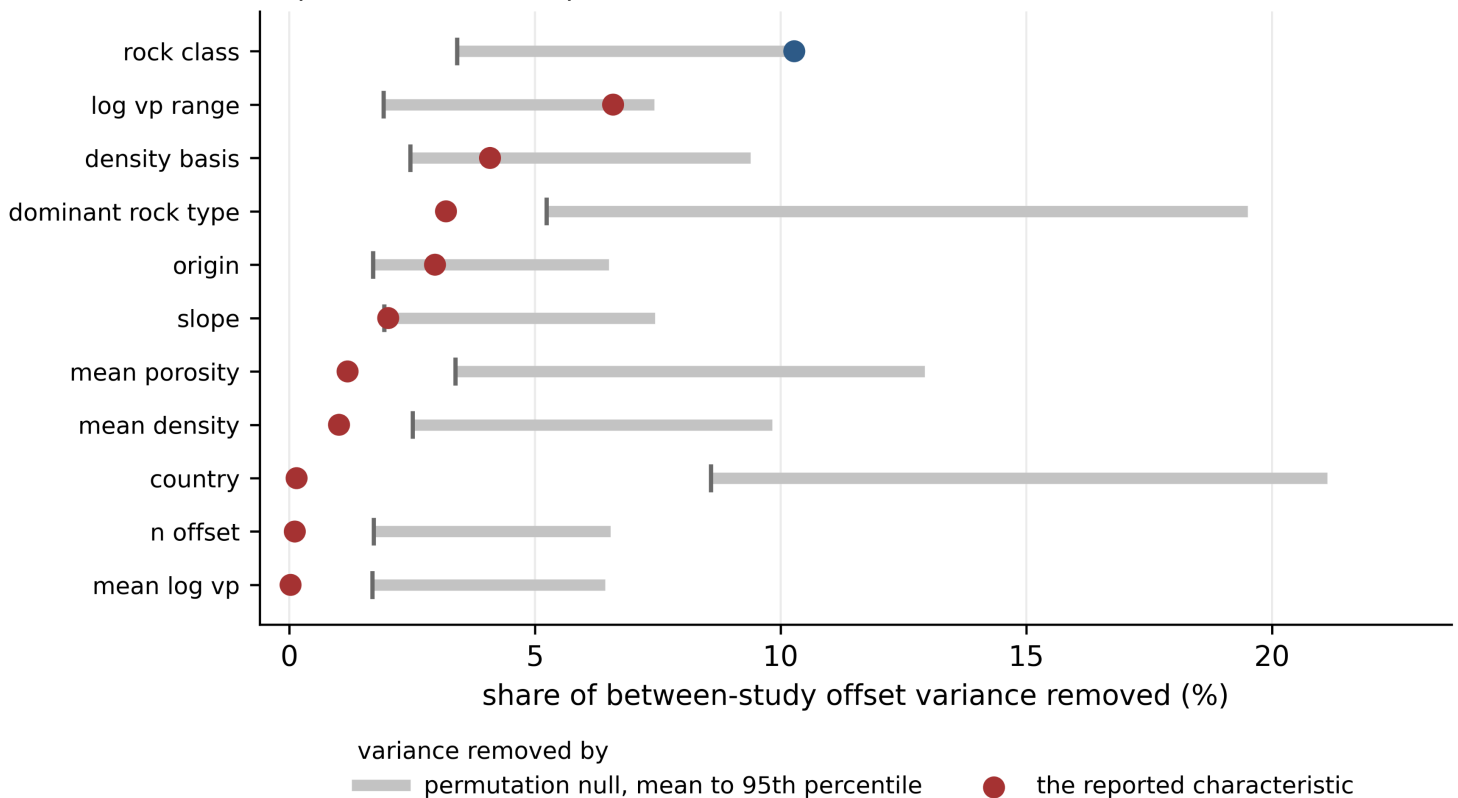


Figure 5

Share of between-study offset variance removed by each reported study characteristic (circles), against the permutation null for the same moderator (bar spanning the null mean to its 95th percentile). Only rock class clears the 95th percentile of its own null, by 0.08 of a percentage point, and it does not survive correction for the eleven moderators tried. Country removes 0.15 percent where a randomly relabelled moderator with as many levels removes 8.58 percent on average.

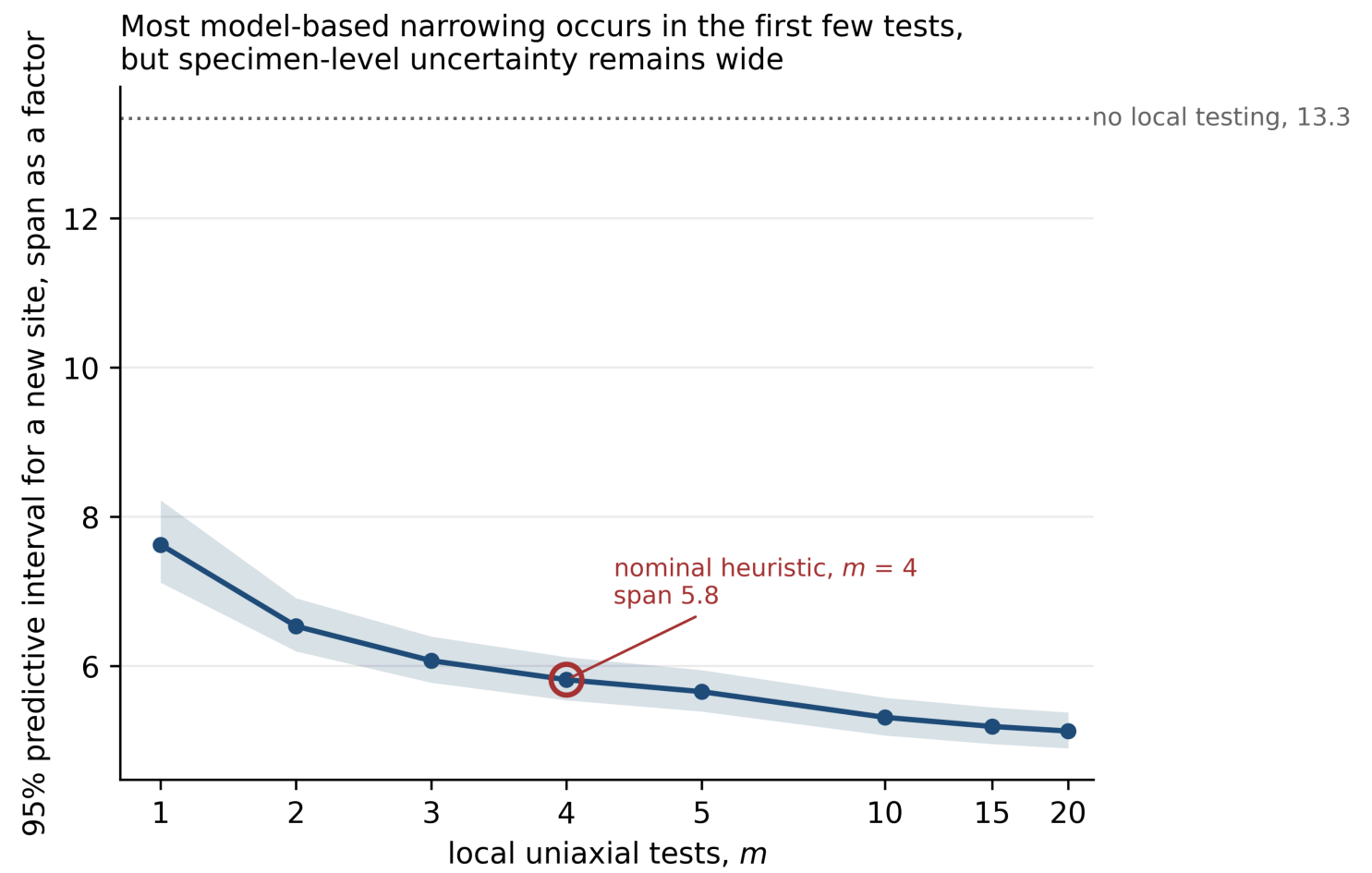


Figure 6

Model-based 95 percent predictive interval for one further specimen from a locally calibrated study population as the number of local UCS tests increases. The band propagates posterior uncertainty in both variance components. The circled point marks the four-test value selected by the nominal marginal-gain heuristic; sensitivity to that heuristic is reported in the text.

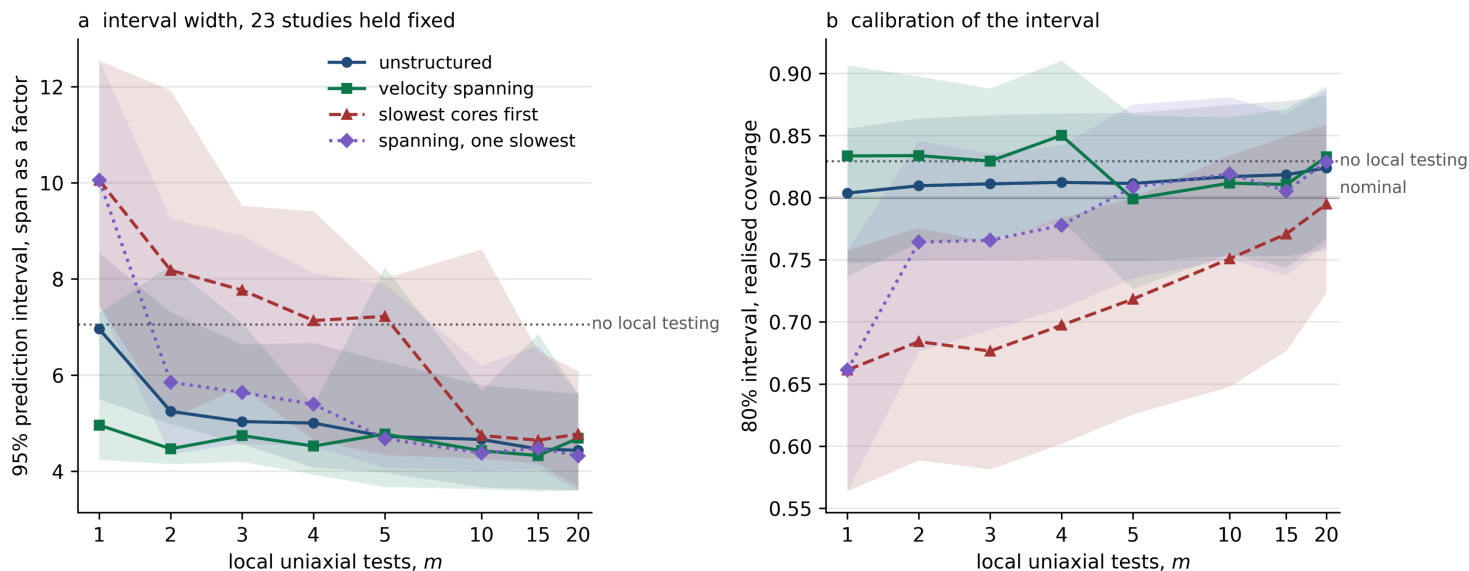


Figure 7

Held-out local calibration on the fixed cohort of 23 studies. (a) Median span of the 95 percent prediction interval. (b) Realised coverage of the nominal 80 percent interval, with the no-testing reference marked. Bands are 95 percent cluster-bootstrap intervals. Unstructured selection remains near nominal coverage, while slowest-first selection under-covers across the tested calibration sizes.

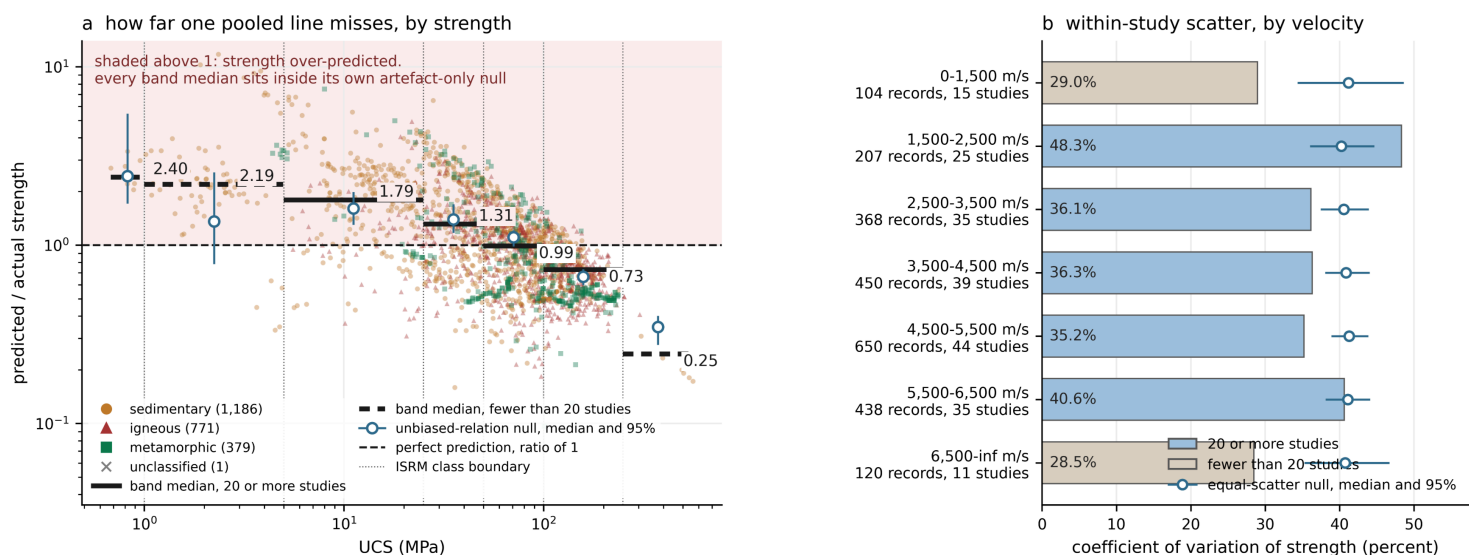


Figure 8

Weak-rock diagnostics for the pooled power relation. (a) Prediction-to-observation ratios grouped by observed strength, with medians from an exactly unbiased artefact-only null. (b) Within-study residual coefficient of variation grouped by velocity, which avoids defining groups with the outcome. Light shading marks bands represented by fewer than 20 studies.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryData1AllCandidateRecordsandScreening.csv](#)
- [Listofsymbols.docx](#)