Estimation of sports volunteering rates in English local authority areas using the Sports England ActiveLives survey

James Liley, Bilal Ashraf, Caroline Dodd-Reynolds, Jochen Einbeck

Short-form summary

Introduction

In this work, we aim to estimate frequencies at which individuals aged above 16 years in English local authority areas volunteered in sport activities, using the Sports England ActiveLives survey. Although the survey includes questions on whether an individual volunteered in the past month and past year, missingness in these responses is non-random, dependent on local authority area and other responses which are also associated with Yes/No answers. We aim to impute these missing values, allowing unbiased estimates of volunteering rates to be made.

Methods

We fitted a logistic linear model and evaluated its accuracy in predicting volunteering status in the past month and year. We explored univariate and multivariate associations with volunteering status, in particular using a topic model to identify patterns in survey answers and examining associations with these patterns.

Results

Volunteering status in past month and past year could be accurately predicted (AUROC>0.78) with good calibration. Estimates of volunteering rates and errors for the English population revealed differences across local authority areas in volunteering rates. We found a range of differences amongst individuals who answered ‘No’ and ‘Yes’ to .

Conclusions

We conclude that volunteering rates differ by local authority area in England. We identify a range of predictors of volunteering status, and describe their relative importance.

Overview

In this work, we document an analysis of the ActiveLives survey by Sport England with the dual aim of identifying factors associated with whether an individual engages in sport volunteering and inferring rates of sport volunteering in English local authority areas

To evaluate sport volunteering activity we consider two responses from the ActiveLives survey essentially corresponding to whether a given individual engaged in sport volunteering in the past month and past year respectively. Both of these survey responses have substantial not-at-random missingness, hampering direct estimation of average responses by local authority area. Our overall strategy is to construct a predictive model for each response which can be used to impute missing values, thereby improving estimates of local authority area volunteering rates. We use an L-1 penalised logistic regression model (LASSO) for prediction, where the L-1 penalty serves both to improve predictive accuracy and to improve model interpretability.

We evaluate associations with volunteering status in three ways. Firstly, we estimate univariate association between individual predictors and volunteering status. Secondly, we make estimates of the coefficients of various other survey responses in a logistic model to predict volunteering status. Thirdly, we use a topic model to identify commonly-associated sets of responses and evaluate association of such topics with volunteering status.

This document is organised as follows. Firstly, we detail the raw data and our overall data handling strategy and describe data processing. We then give an initial analysis of volunteering rates by local authority area and explore univariate associations. We then describe our predictive models and how these improve estimation of local-area rates. Finally, we explore abstract associations using a topic model.

Throughout, the term ‘volunteering status’ will be used to mean either volunteering status in the past month or in the past year, depending on context, and may refer to either a reported answer (‘yes’, ‘no’, or ‘NA’) or an imputed value (a probability that the true answer is ‘yes’). The term ‘volunteering rate’ will be used to indicate a mean volunteering rate in a population.

1. Details of raw data

The raw dataset contains 187533 observations of 7550 variables. Rows are identified with a serial number (‘serial’) and a recommended demographic weighting ‘wt_final’ is associated with each sample. Variables include

- Circumstances of the survey (online or in person, time e.g. ‘Type’, ‘group’, ‘month’)
- Demographics (age, sex, ethnicity, children e.g. ‘Age1640’, ‘Gend3’, ‘Eth4’, ‘Child4’)
- Many filters (e.g. ‘Filter_Act’, ‘Filter_Act_F’)
- Other weightings (e.g. ‘wt_final_AB’, ‘wt_time’)

- All survey-specific variables

Of particular concern in our analysis are the following response variables:

- Volint_ANY: 'Whether done any volunteering in last 12 months'
- VOLMTH_POP: 'Volunteered in the last 4 weeks excluding those doing solely raising funds (REBASED)'

We are concerned with predicting volunteering rates per local authority area, with the latter being defined according to the variable 'LA_2021' (LA 2021 codes). Throughout, we use weights 'wt_final' to weight samples.

Missingness of important variables is summarised in table 1:

Variable	**Description (short)**	**Missing (%)**	**Non-missing (%)**
Volint_ANY	Volunteered in last year	63300 (34)	124233 (66)
VOLMTH_POP	Volunteered in last month	76603 (41)	110930 (59)
wt_final	Demographic weighting	1788 (1)	185745 (99)
LA_2021	local authority area	0 (0)	187533 (100)

Table 1: Missingness in key variables

The majority of individuals with missing weights were under 16 (1675 of 1788, 94%). The number of over-16 individuals with missing weights was small (113), so such individuals were removed along with under-16 individuals in all subsequent analyses.

2. Protocol for train/test/validation split

Note that only data with non-missing target values (volunteering status) could be used for training or testing the model. Our eventual use of the model was to generate predictions for samples for which volunteering status was missing.

Central to our protocol is the assumption that our target (volunteering status in last year or month) is conditionally independent of target missingness status (whether the volunteering status question was not answered) conditional on all other variables (predictors). This is an unavoidable assumption, as it is effectively stating that the target can be predicted when it is missing by making inferences about the behaviour of the target from predictors when the target is non-missing.

We divided data randomly by samples into ‘training’, ‘validation’ and ‘test’ sets in the ratio 3:1:1. Since we eventually estimated the accuracy of our predictor on the test set, we could not use performance on the test set to judge which of several potential candidate models to use. We thus made decisions around model choices (e.g., include PCAs or not, include location as a random effect, etc) by training candidate models to the training set only and evaluating their performance on the validation set (figure 1, leftmost).

LASSO models roughly select the ‘best-performing’ predictors only to include in a final model. This means that the coefficient estimates of these best-performing predictors in the LASSO model may be higher than they would be in an independent dataset (e.g., they overestimate the real effect) because of regression-to-the-mean effects (note that they are not formally biased as coefficients in a logistic regression model do not have finite expected values). In order to make accurate estimates of coefficients, we firstly fitted L-1 penalised models to the training set to choose variables, then estimated the coefficients of these variables on the aggregated test and validation sets (figure 2, column 2).

To evaluate the final predictive performance, we fitted a LASSO model to the aggregated training and validation sets and evaluated it on the test set (figure 1, column 3)

We fitted the final model to perform imputation on volunteering rates using all non-missing data (figure 1, column 4). We fitted our topic model to the training set and evaluated posterior probabilities of topic membership on the test and validation sets.

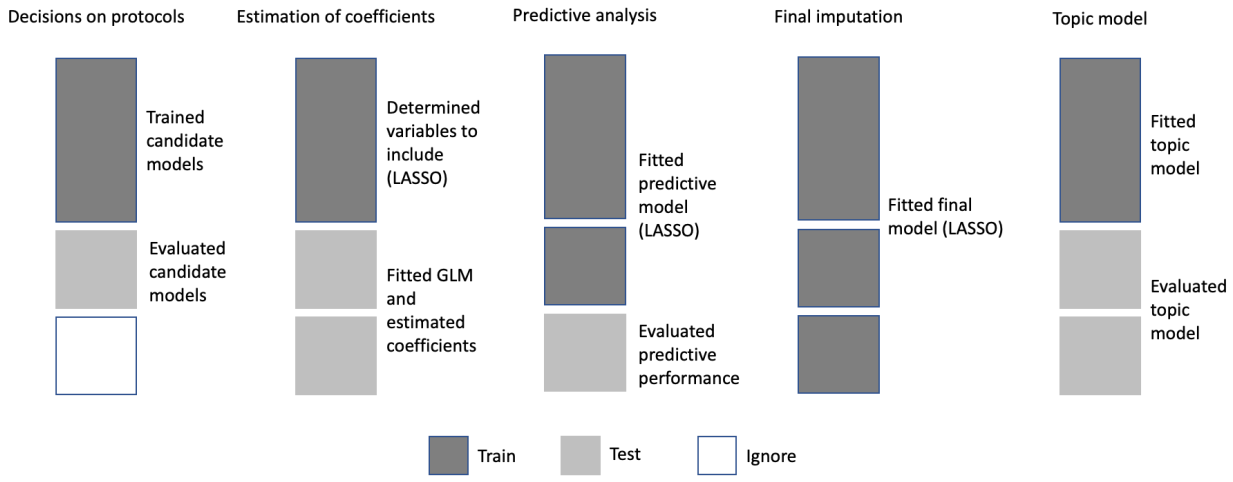

Figure 1. Training/test/validation scheme

This document and testing protocol was reviewed by all authors before any data in the test set was viewed.

Use of L-1 penalised regression methods (LASSO) requires specification of a hyperparameter λ governing the extent of penalisation. Whenever we used a LASSO model, we first tuned this hyperparameter using ten-fold cross validation on the same dataset to which the model was being fitted.

3. Initial analysis of volunteering rates by local authority area and other variables

Note on populations and estimates

An important point in this analysis is consideration of the potential populations we can analyse. We may be interested in volunteering rates and other characteristics associated with volunteering rates in any of the four populations:

1. The population of individuals who, if asked to complete the survey and asked about volunteering, would do the survey and answer the volunteering question.
2. The population of individuals who, if asked to complete the survey, would do so
3. The whole English population.
4. The population of individuals who, if asked to complete the survey and asked about volunteering, would do the survey but **not** answer the volunteering question.

The population and proportion in 1 is easy to study, as the set of people in the survey who answered the volunteering question is essentially a random sample from this population. We can readily find associations with volunteering in this population by simply analysing our data directly.

The population and proportion in 2 is harder to study, as it includes people who did the survey but did not answer the question. We establish in the subsequent section that such missingness in this question is not at random, and we cannot assume that none of the individuals who did not answer the question did not in fact volunteer, so the proportion of 'yes' answers to 'yes'/'no' answers is not appropriate as an estimator of volunteering rates in this population. By predicting answers to the volunteering question, we can make an estimate of volunteering rates in this population.

The population and proportion in 3 is the most important but the hardest to study. We may attempt to estimate proportions of individuals who volunteered in this population by demographic weighting. The weight assigned to an individual (column wt_final) is the ratio of the frequency of their demographic values in the survey to the frequency of their demographic values in the general English population. In order for our weighted mean of reported and imputed volunteering status estimates to be unbiased, we require an assumption that the chance of an individual agreeing to do the ActiveLives survey is conditionally independent of predictor variables given weighting. That is, if we have ten individuals with demographic weight

0.3 but with different demographics, e.g. different age, the chance of any of them saying 'yes' to doing the survey is the same and does not depend on age. We also require an assumption that for any given weight, our mean imputed volunteering estimate amongst individuals of that weight is unbiased.

Final weights are computed with respect to the entire English population rather than on a local-area basis. Hence, in order for weighted means of volunteering statuses to give unbiased estimates of volunteering rates in population 3 in local authority areas, we must make the additional assumption that the relative frequency of a demographic class in a local authority area compared to England is the same in the survey population and in the general population.

Finally, we may make an unbiased estimate of the proportion in population 4 by taking an unweighted mean of imputed volunteering status amongst individuals for whom reported volunteering status was NA (not answered).

Analysis by local authority area

We first counted the number of ‘Yes’, ‘No’ and ‘Missing’ answers to volunteering questions (VOLMTH_POP: last month and Volint_ANY: last year) for each of the 309 local authority areas. Full data is shown in supplementary table 1, amongst columns marked ‘Raw data’. We computed both raw means (column Raw_pct_yes) of volunteering status (number of individuals volunteering over number answering) and means weighted by ‘wt_final’ (column Raw_wt_pct_yes). We also computed raw (Raw_pct_na) and weighted (Raw_wt_pct_na) means of the proportion of answers that were NA amongst individuals from each local authority area. In terms of the populations mentioned in the previous subsection, the columns ‘Raw_pct_yes’ are estimates of volunteering rates in each local authority area in population 1. We cannot make estimates for populations 2 and 3 at this stage. Throughout this section, we rejected null hypotheses for p-values less than 0.006, in order to control FWER at <5% over eight independent tests.

We firstly investigated whether missingness rates in Volint_ANY (MY; ‘Missing-Year’) and VOLMTH_POP (MM; ‘Missing-Month’) differed between LA_2021 categories (LA). Using an unweighted test of equality of proportions against null hypotheses that MY and MM were independent of LA, we rejected the null hypotheses for Volint_ANY ($p<1 \times 10^{-8}$) but not for VOLMTH_POP ($p=0.027$). We conclude that missingness in the answer to Volint_ANY depends on local authority area. An implication for the populations mentioned in the previous section is that estimates of volunteering rates in population 1 are likely to be biased as estimates of volunteering rates in population 2.

Since distributions of demographic weightings (WT) differ substantially across LAs, we further investigated whether MY or MM were conditionally independent of LA given WT; that is, whether local authority area has an additional effect on MY or MM amongst people of the same WT. Using nested logistic linear regression models with a likelihood ratio test, we rejected the null

hypothesis of conditional independence of MY and LA given WT ($p=5 \times 10^{-8}$) but did not reject the null hypothesis of conditional independence of MM and LA given WT (p=0.02). We conclude that local authority area affects missingness in answers to Volint_ANY independently of effects from weight. This incentivises our project, as it suggests that missingness in volunteering rate answers is not at random.

We tested whether the proportion of individuals answering 'Yes', as a proportion of individuals who gave an answer, differed between local authority areas. Against a null hypothesis of equal Yes/No proportions in each local authority area, we rejected the null hypothesis for Volint_ANY and for VOLMTH_POP ($p<1 \times 10^{-10}$ for both) . We conclude that the ratio of number of individuals answering 'Yes' to number of individuals answering 'No' differs across local authority areas (this relates to population 1 above)

Univariate predictors of volunteering rates

Individuals who answered 'Yes' were different to individuals who answered 'No' across many survey variables, but in many cases also had different missingness rates in the survey variable. We considered a range of variables intended to be roughly representative of demographics. We took a robust association of volunteering status with some variable as an association with that variable ($p < 0.0005$, to control FWER at < 5%, given approximately 100 independent tests) and non-association with missingness in that variable ($p > 0.05$). For association tests with binary variables (including missingness) we used a binomial test of proportions, for categorical variables we used a chi-squared test of multiple proportions, and for ordinal variables we used a Wilcoxon signed-rank test. Only non-missing values of volunteering status were used; hence these are associations for population 1 only (see supplementary tables 1 and 2). Samples with missing values of the other variable were removed if missingness was not associated with missingness in volunteering status.

The robust differences between individuals who answered 'Yes' and individuals who answered 'No' to Volint_ANY were number of attendances of a live sport event in the prior 12 months (more Yes if once or more), number of children in household (more Yes if one or more), ONS groups (more Yes for farming communities, rural tenants, ageing rural dwellers, students around campus , urban professionals and families, ageing urban living, suburban achievers and semi-detached suburbia); ONS super-groups (more Yes in rural residents, urbanites, suburbanites); IMD deciles (more Yes in less deprived deciles); gender (more Yes for male and other); RUC 2011 (more Yes for rural); community sports partnership (more Yes for Buckinghamshire and Milton Keynes, Peterborough & Cambridgeshire, Cheshire and others); mode of completion (more Yes for online); and month of completion (more Yes for May-Nov).

The robust differences between individuals who answered 'Yes' and individuals who answered 'No' to VOLMTH_POP were similar: mode of completion (more Yes when online), highest qualification level (more Yes when higher), gender (more Yes when male), interview month

(more in May-Aug, and Nov-Feb), IMD deciles (more Yes when less deprived), ONS super-groups (more Yes for rural residents, urbanites and suburbanites), ONS segments (more Yes in farming communities, rural tenants, ageing rural dwellers, students around campus, urban professionals and families, ageing urban living, suburban achievers, semi-detached suburbia), community sports partnership (more Yes in Buckinghamshire and Milton Keynes, Peterborough & Cambridgeshire, Cornwall and Isles of Scilly, and others) and RUC 2011 (more Yes in rural).

All other variables under consideration were removed due to missingness being associated with Volint_ANY/VOLMTH_POP; it is likely that more associations would have been found were data more complete. See supplementary table 3 for more details

4. Data cleaning and processing for predictive model

For our predictive analysis, we performed the following data processing steps, in order:

1. Filtered to samples over 18, using the variable Age16plus
2. Filtered to
   - Non-missing weight (wt_final)
   - Non-missing target (Volint_ANY for previous-year analysis, VOLMTH_POP for previous-month analysis)
3. Removed several variables
   - Relating to volunteering (containing ‘vol’ or ‘Vol’, and ‘VOLFRQ_POP’),
   - Filters (containing ‘Filter’),
   - Weighting variables other than final weights (containing ‘wt_’, excepting ‘wt_final’)
   - local authority area variables (containing ‘LA_’)
4. Removed any variable with missingness >50%
5. Converted any variables for which answers were ‘Yes/No’ such that ‘Yes’ always corresponds to 1, and ‘No’ to 0
6. Mean-value imputed missing values for all variables other than
   - Weights (wt_final),
   - Target (Volint_ANY for previous-year analysis, VOLMTH_POP for previous-month analysis)
7. Removed variables with low variation: standard deviation after mean value imputing ≤1 x 10^{-5}
8. Normalised all variables except
   - Weights (wt_final),
   - Target (Volint_ANY for previous-year analysis, VOLMTH_POP for previous-month analysis),

   to mean 0 and standard deviation 1.

Whenever a model fitted on a set of data A was used to make predictions on a set of data B, all imputation, normalisation and exclusion rules (steps 4, 6, 7 and 8 above) were determined on set A only. For example, for candidate models (leftmost column in figure 1), all imputation and

normalisation rules were determined on the training set, and variables were removed if they had missingness >50% or post-imputation standard deviation ≤ 1×10^{-5} amongst values in the training set. When imputing volunteering status variables (Volint_ANY, VOLMTH_POP), only survey samples for which these variables were non-missing were used to determine imputation, normalisation and exclusion rules.

It is possible that our mean-value imputation procedures are slightly biased, in that some variables may not be missing-at-random. However, given the low-sensitivity character of the survey material, we expect that deviation from missingness-at-random is small and mean-value imputation is justified.

We also considered some feature engineering, using the protocol in the leftmost column of figure 1:

- Addition of projections to principal components of most sparse variables
- Addition of posterior probabilities of topic membership for fifty topics
- Inclusion of a measure of local authority area (LA_2021) as a random effect

We found that these made negligible or undetectable differences to predictive power, and left them out in the interests of simplicity.

5. Prediction of volunteering rates using per-month responses

We report two analyses: firstly, a model fitted to our training set used to estimate coefficients of predictors in a logistic linear model and secondly a model fitted to our training and validation sets to estimate predictive performance (see figure 1, columns 2 and 3). Since the sets of samples used to fit the model differ, the set of variables removed and the set of variables included in the LASSO model differ slightly between the two analyses.

Firstly, we processed the training set. After processing, the dataset for training consisted of 54488 observations of 2438 predictors. We then processed the validation and test sets using the imputation, normalisation and exclusion rules determined by the training set, after which this combined dataset contained 43727 observations. We fitted an L-1 penalised linear model to the training set only, and tuned the L-1 penalty parameter λ using ten-fold cross-validation on the training set.

We then fitted a generalised linear model to the combined test and validation sets including only the variables included in the LASSO model fitted to the training set. Coefficients and associated p-values (against the null hypothesis that inclusion of the associated variable in the linear model does not improve prediction independently of other variables) are shown in supplementary table 4. It should be noted that since mean-value imputation is used, informative missingness in predictors may contribute to association. Amongst the variables with significant coefficients were number of times in past 12 months attended a live sports event (more Yes with more often and non missing), number of children (more Yes with more children and non missing), cheerleading

1-3 months before (more Yes if yes), and team sport per week (more Yes with more time). More details of associated variables are given in table 4; only the test and validation datasets are used to find frequencies for this table.

We then processed the test set using the imputation, normalisation and exclusion rules determined by the combined training and validation set. After processing, this left 2444 predictors with 87525 observations for training and 21790 for testing. We then fitted an L-1 penalised linear model to the combined training and validation set, and evaluated its performance on the test set.

We visualised performance on the test set using ROC/PRC curves and calibration curves, shown in figure 2. The AUROC was 0.892 (SE 0.007) and AUPRC was 0.280 (SE 0.015). The AUROC may be interpreted as: for a random individual who answered 'Yes' and a random individual who answered 'No', the probability that the first individual was assigned a higher probability of volunteering in our model was 89.2%. The predictive model was well-calibrated at low probabilities of volunteering, and less well-calibrated at probabilities >0.5; however, most individuals had a predicted probability of volunteering well under 0.5 (see black dots on figure).

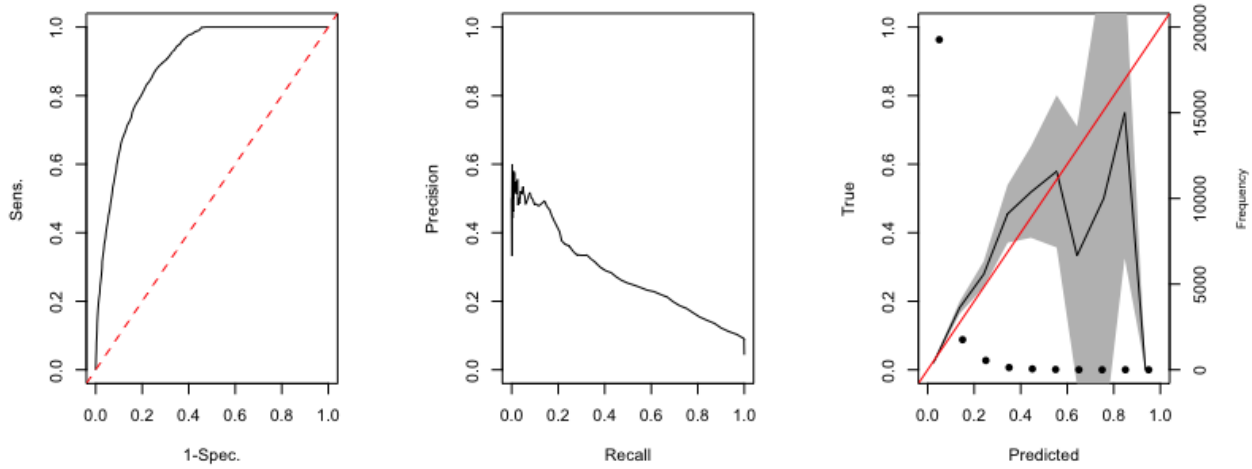

Figure 2. ROC, PRC and calibration curves for predictor for VOLMTH_POP (volunteered in past month). Grey region on calibration plot is a 95% pointwise confidence interval, using bins of width 0.1. Black points on calibration plot correspond to right axis and indicate the frequency of predicted y-values in bands of width 0.1.

We evaluated calibration of volunteering rate estimates in each local authority area. For each local authority area, we computed and compared the frequency of 'Yes' answers to VOLMTH_POP amongst non-missing answers to VOLMTH_POP and the mean predicted probability of a 'Yes' answer for individuals in the test set. We then ran a binomial test for deviation of the two values for each region against the null hypothesis that the true frequency of 'Yes' answers corresponded to the mean predicted probability. No region showed significant

deviation between predicted and observed frequency (min p-value 0.0145; Sidak-corrected significance threshold to control FWER < 5% given 309 independent tests 0.00017), and we conclude that our predictions are consistent with observed frequencies of VOLMTH_POP in all local authority areas. Results are shown in supplementary table 5.

We then reconstructed volunteering rates in each local authority area. We estimate mean frequency of last-month volunteering in each population as follows:

- For population 1, using unweighted frequencies of 'Yes' responses amongst all non-missing responses (columns Raw_pct_yes and Raw_pct_yes_se)
- For population 2, we may make unbiased estimates by combining unweighted means of responses across individuals who answered the VOLMTH_POP question and mean predicted probabilities of answering 'Yes' amongst those who did not (columns Final_pct_yes and Final_pct_yes_se)
- For population 3, we make unbiased estimates of volunteering rates by local authority area as for population 2 but using weighted means (columns Final_wt_pct_yes and Final_wt_pct_yes_se)
- For population 4, we make unbiased estimates as unweighted means of predicted values across individuals with missing values for VOLMTH_POP (columns Imp_pct_yes and Imp_pct_yes_se).

Standard errors of weighted means were estimated using bootstrap resampling. Error in prediction was small for most estimates (see calibration plot; note most predictions are less than 0.3 or 30%). All estimates are shown in supplementary table 1.

6. Prediction of volunteering rates using per-year responses

We analysed per-year volunteering status (question Volint_ANY) in the same way as per-month volunteering status.

After processing, the dataset for training consisted of 73377 observations of 2438 predictors. The combined test and validation dataset contained 49081 observations. Coefficients and association tests for the linear model fitted to the combined test and validation datasets for variables included in the LASSO model fitted to the training dataset are shown in supplementary table 4. Amongst variables with significant coefficients were number of times attended a live sports event in the past 12 months (more Yes with more often), number of children (more Yes with more), highest qualification level (more Yes with higher), motivation for sport/exercise not want to disappoint other people (more Yes if agree), age 16-40 (more Yes in age 16-40), motivation sport/exercise enjoyable and satisfying (more Yes if agree), number of adults in household (more Yes with more), whether 'if I find something difficult, I keep trying until I can do it' (more Yes with disagree), limiting disability (more Yes with non-limiting or no disability), and others. More details of associated variables are given in supplementary table 4.

For the analysis of predictive performance, after processing, the combined training and validation sets had 2444 predictors with 97977 observations for training and the test set had 24481 observations for testing.

ROC/PRC curves and a calibration curve are shown in figure 3. The AUROC was 0.784 (SE 0.004) and AUPRC was 0.502 (SE 0.007) (with the interpretation that: for a random individual who answered 'Yes' and a random individual who answered 'No', the probability that the first individual was assigned a higher probability of volunteering in our model was 78.4%). The predictive model was well-calibrated across all probabilities.

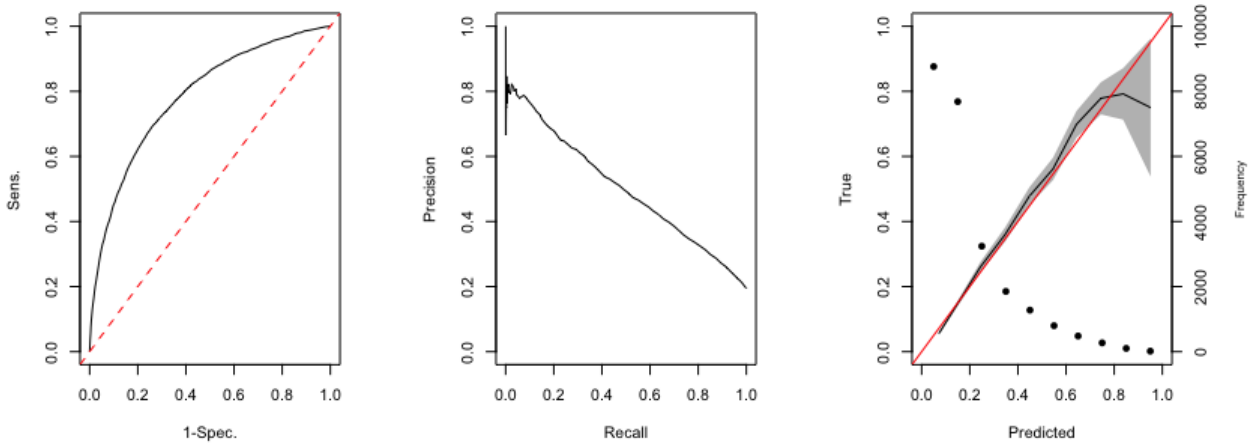

Figure 2. ROC, PRC and calibration curves for predictor for Volint_ANY (volunteered in past year). Grey region on calibration plot is a 95% pointwise confidence interval, using bins of width 0.1. Black points on calibration plot correspond to right axis and indicate the frequency of predicted y-values in bands of width 0.1.

We again evaluated calibration of volunteering rate estimates in each local authority area, and again no region showed significant deviation between predicted and observed frequency (min p-value 0.0055; Sidak-corrected significance threshold to control FWER < 5% given 309 independent tests 0.00017), and we conclude that our predictions are consistent with observed frequencies of Volint_ANY in all local authority areas (see supplementary table 5).

We reconstructed volunteering rates in each local authority area as per the previous section. Full output is shown in supplementary table 2.

7. Details of fitted models

For each prediction task, we fitted three LASSO models: one to identify variables for which coefficients should be estimated, one to evaluate predictive accuracy, and one for final imputation. These models were fitted in similar circumstances but used different volumes of training data. The L_1 hyperparameter governs the degree of coefficient 'shrinkage' in the model,

with a higher value corresponding to fewer variables included. Details of models are given in tables 2 and 3 below.

	N training obs.	L_1 hyperparam.	N predictors	N inc. predictors
Coef. estimation	65588	0.00231	2438	212
Predictive analysis	87525	0.00192	2444	248
Final	109315	0.00178	2456	260

Table 2: details of LASSO models fitted to predict VOLMTH_POP (previous-month volunteering rates). Columns are as for table 2.

	N training obs.	L_1 hyperparam.	N predictors	N inc. predictors
Coef. estimation	73377	0.00258	2438	321
Predictive analysis	97977	0.0023	2444	336
Final	122458	0.00209	2456	370

Table 3: details of LASSO models fitted to predict Volint_ANY (previous-year volunteering rates). Column N_training_observations is the number of observations in the training set for the model. Column 'L1 hyperparam.' is the value of the L1 penalty λ in the LASSO model (determined by ten-fold cross validation). Column 'N inc. predictors' is the number of predictors with nonzero coefficients in the fitted model.

8. Further analysis of variable associations with volunteering rates using topic models

Topic models use a statistical method called Latent Dirichlet Allocation to find sets of survey responses which typically occur together, and provide a useful clustering method giving insight into typical groups of survey responders. We processed our training set in the same way as for the predictive analysis. We then restricted only to the variables beginning with any of: ACTY, DAYS, DUR_, DURA, FREQ, HABI, INOU, MEMS, MINS, MONT, SETI, SETO, SURF (general survey variables) and added the following variables:

- Male sex (Male) and female sex (Female)
- Under 40 (U40) and over 40 (O40) X2$Male=(X$Gend3==1)

- Index of multiple deprivation (IMD) greater than 5 (High_IMD) or less than or equal to 5 (Low_IMD)
- Any child under 13 (ChildU13) or not (NoChildU13)
- Qualifications level 3 and above (Qual_34) or not (Qual_012)
- Limiting disability (LimitingDisability) or not (NoLimitingDisability)
- In full or part time work (Working)

We interpreted each variable as ‘yes’/’no’ (taking 0 to mean ‘no’ and >0 to mean ‘yes’, rounding imputed values). We then fitted a topic model with 50 topics. For each topic, we identified the variables for which the posterior probability of inclusion in that topic was greater than 1% (noting that posterior probabilities over all variables sum to 1). The set of variables associated with each topic is shown in supplementary table 5. We assigned labels to topics where obvious; for instance, topic ‘Adventure’’ contained FREQUENCYGR_ADVENTURE_D01, DAYS10P60GR_HILLWALK_R03, DUR_HVY_ADVWATERSPORT_C07, amongst others.

We then processed the combined validation and test sets using the imputation, normalisation and exclusion rules determined by the combined training set. We then computed the posterior distribution of each sample over topics, and for each topic assessed whether individuals who answered ‘Yes’ to Volint_ANY or VOLMTH_POP had significantly different probabilities of being associated with that topic than did individuals who answered ‘No’.

A topic model can be thought of as finding ‘groups’ in the population defined by patterns of answers, and the posterior as assigning each individual to ‘groups’ to varying degrees, so that group assignment degrees for each individual sum to 100%. We compared distributions of posterior topic probabilities for individuals answering ‘yes’ or ‘no’ to Volint_ANY or VOLMTH_POP using Wilcoxon signed-rank tests. Because a Wilcoxon-signed rank test compares probabilities of random values from each set being larger or smaller than the other, we first rounding posterior topic probabilities to the nearest percent, reasoning that differences in topic probabilities of fractions of a percent were not meaningful compared to differences of >1%. We rejected null hypotheses for a p-value threshold of 0.0005 to control FWER at <0.05 for 100 independent tests.

Individuals who answered ‘Yes’ to Volint_ANY (volunteered in past year) had significantly higher topic probabilities for topics "Combat", "Cycling", "Cycling_travel", "Dance_golf", "Football" "Rackets", "Running", "Swimming", and "Watersport", and significantly lower topic probabilities for "Walking", "Walking_active", "Walking_leisure", "Gym_cycling", "Gym_general", "Equestrian" "Dance" and "Gardening".

Individuals who answered ‘Yes’ to VOLMTH_POP (volunteered in past month) had significantly higher topic probabilities for topics "Running", "Watersport", "Winter_sport", "Leisure", "Cycling_travel", "Rackets", "Dance", "Intervals", "Cycling", ”Football", and “Adventure”, and significantly lower topic probabilities for "Gym_class", "Gym_cycling", "Gym_weights", and "Equestrian". Topics corresponding to ‘Walking’ and ‘Gym_general’ were variably overrepresented amongst ‘Yes’ and ‘No’ answerers.

Full results are shown in supplementary table 5.

Supplementary tables

Please see
https://docs.google.com/spreadsheets/d/1fahd-NthidaUGeT0R5qwybO6XZDyzaAkTbFaInLFxP8/edit?usp=sharing