Supplementary information

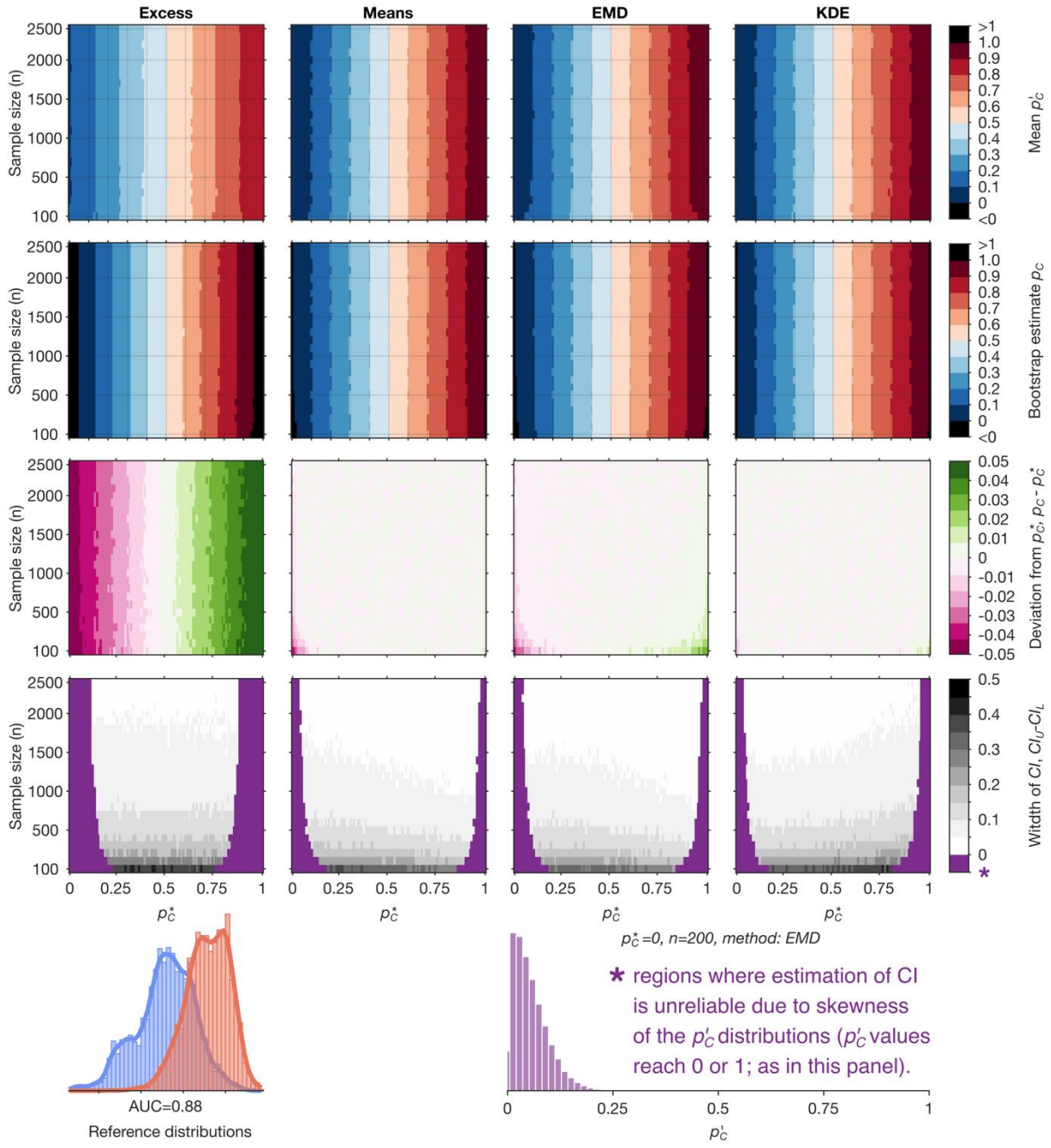
Supplementary Text and Figures

Quantitative characterisation of methods and generation of heatmaps

To systematically compare all the methods. We stochastically, modelled mixture populations comprising of random samples (sampled from the reference populations with replacement) at defined proportions of each reference population, p_C^* and sample size, n. The proportion and sample size were systematically varied, with p_C^* ranging from 0 to 1 in 0.01 (1%) steps while n ranged from 100 to 2,500 in steps of 100 samples. All four methods were applied to each combination of these parameters. At each point in the parameter space, we estimated the prevalence (p_C) and its confidence intervals (following the methodology from the paper) and we compared it with the model proportion (p_C^*) used to generate them. This idealised scenario allows a direct head-to-head comparison of accuracy between all four methods. Results of this comparison are presented in Supplementary Figs 1 and 2.

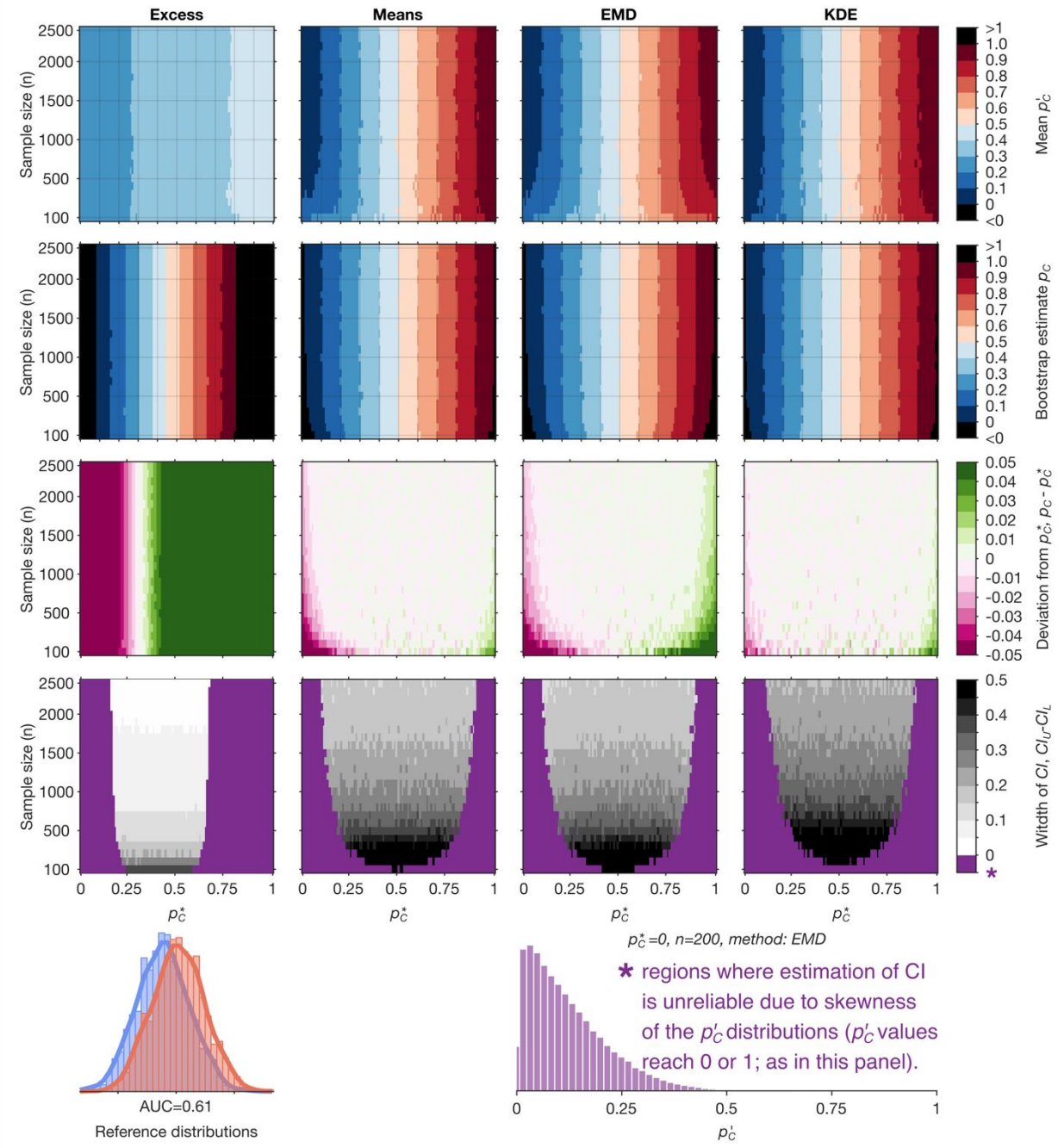
Supplementary Figs 1 and 2 show:

- (1st row) Mean value of the estimates of prevalence $\boldsymbol{p_C'}$ in the 100,000 bootstrap samples,
- (2nd row) Bias corrected prevalence estimate $\boldsymbol{p_C}$,
- (3rd row) Deviation from the true proportion $\boldsymbol{p_C - p_C^*}$,
- (4th row) The width of confidence ($\boldsymbol{CI_U - CI_L}$) intervals,
- Reference populations and an example of a skewed distribution of the $\boldsymbol{p_C'}$ values.



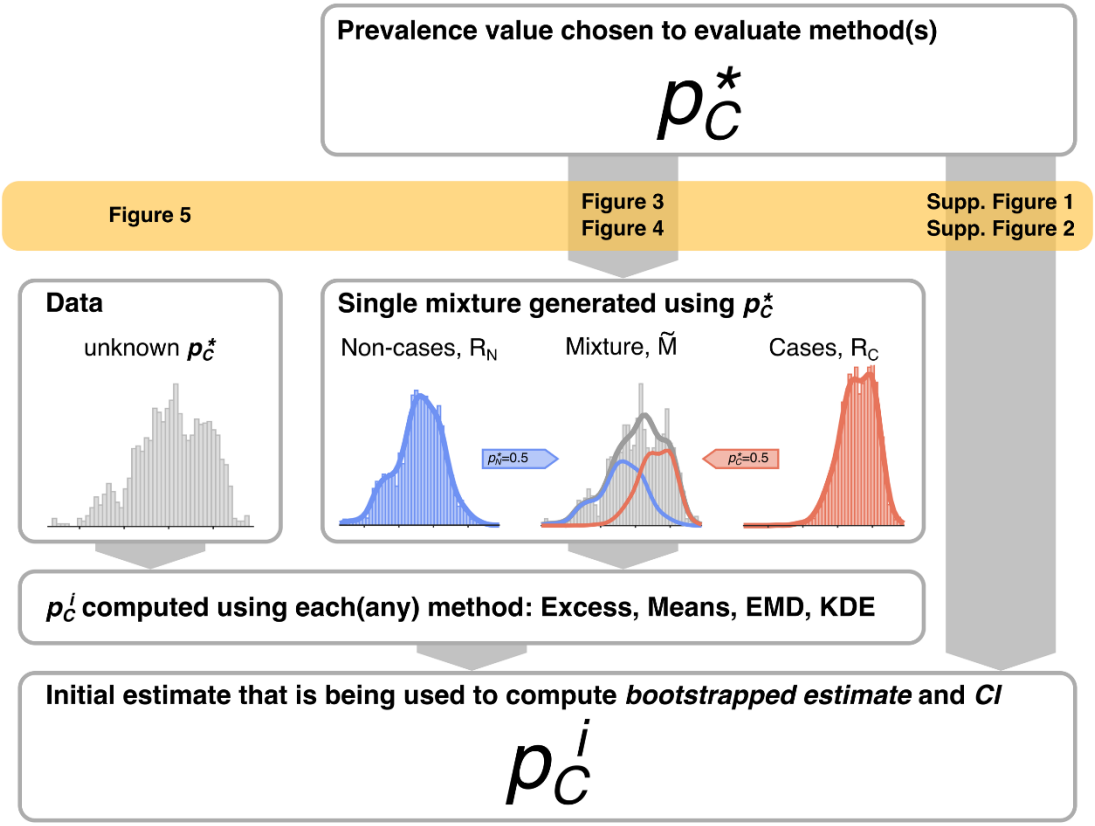
Supplementary Figure 1: A comparison of the four methods using the (Type 1 GRS) dataset. (Top row) Mean value of the 100,000 estimates of prevalence (p'_C) in the bootstrap samples across defined mixture proportions (p^*_C) and the mixture sample size (n) of the dataset for each method. (Second row) Bias corrected prevalence estimates (p_C) across the constructed samples. (Third row) deviation from the true proportion ($p_C - p^*_C$) across the constructed samples. (Bottom row) The width of confidence ($CI_U - CI_L$) intervals of the individual estimates across the constructed samples. The purple colour indicates regions in which the distributions of p'_C are skewed which may affect estimation of the CI. It can be observed that across the parameter space, the Means, EMD and KDE methods all typically outperform the

Excess method. It is also evident that sample sizes of approximately 1,000 or more are generally sufficient for each method. A further increase of sample sizes would be recommended in order to properly estimate CI of the most extreme proportions. Calculations were based on the following participants: type 1 diabetes cases WTCCC, type 2 diabetes cases WTCCC.

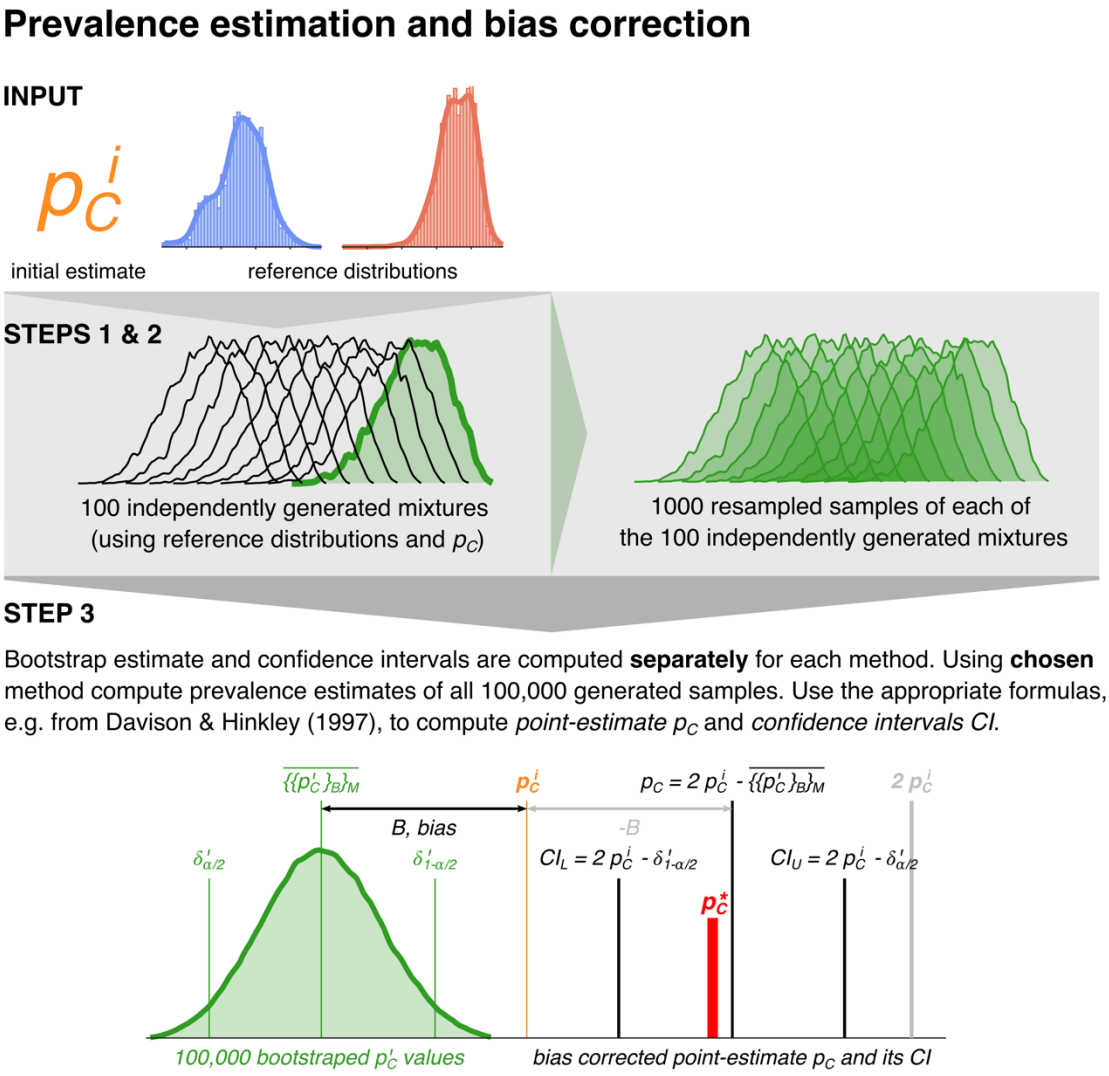


Supplementary Figure 2: A comparison of the four methods using the (Type 2 GRS) dataset. (Top row) Mean value of the 100,000 estimates of prevalence (p'_C) in the bootstrap samples across defined mixture proportions (p^*_C) and the mixture sample size (n) of the dataset for each method. (Second row) Bias corrected prevalence estimates (p_C) across the constructed samples. (Third row) deviation from the true proportion ($p_C - p^*_C$) across the constructed samples. It can be observed that with highly overlapping reference distributions, the Means, EMD and KDE methods all clearly outperform the Excess method. It is also evident that extreme proportions are much harder to estimate for these mixtures compared to the Type 1 GRS mixtures. (Bottom row) The width of confidence intervals ($CI_U - CI_L$) of the individual

estimates across the constructed samples. The purple colour indicates regions in which the distributions of (p'_C) are skewed which may affect estimation of the CI. Estimates with the Type 2 GRS are typically much more variable across the parameter space compared to when using the more discriminative Type 1 GRS scores. Calculations were based on the following participants: type 1 diabetes cases WTCCC, type 2 diabetes cases WTCCC.



Supplementary Figure 3: Illustration of the approaches used to calculate the initial point-estimate p_C^i throughout the paper. The initial point-estimate p_C^i can be estimated from data (real or simulated) or can be fixed by hand. We used simulated data or a fixed value of p_C^i to evaluate the methods by comparing them with the true prevalence p_C^* (results illustrated in Figs 3 and 4, Supplementary Figs 1 and 2, respectively).

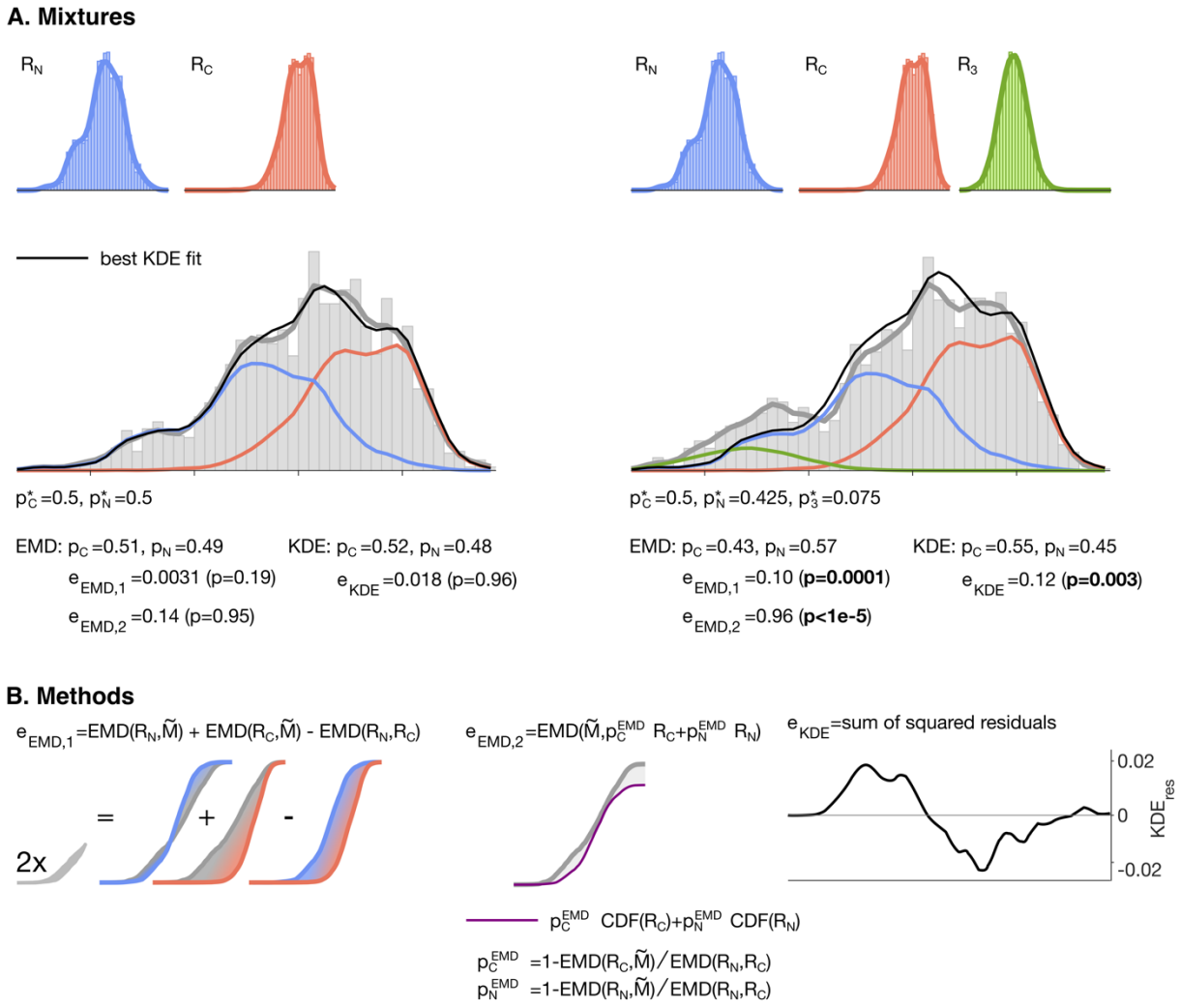


Supplementary Figure 4: Illustration of the steps used in estimation of the prevalence p_C and its confidence intervals. To find a bias corrected estimate of prevalence and its confidence intervals, we are using the initial point-estimate p_C^i and the reference samples. We generate $N_M = 100$ sample mixture populations with a given composition (p_C^i) and sample size (n) equal to the size of the original mixture sample. Next, we resample (with replacement) each of the $N_M = 100$ new mixtures generating $N_B = 1,000$ bootstrap samples. We apply a chosen method to all generated samples and obtain $N_M \cdot N_B = 100,000$ estimates p'_C. We then use the methods described in section "Calculating confidence intervals" to find each bias corrected point-estimate p_C and their confidence intervals. Finally, if we are evaluating the methods, we compare the bias corrected point-estimate p_C with the true prevalence p_C^*. Generally, for real world applications the true prevalence p_C^* is unknown.

The mixture assumption

To check if the mixture assumption, $p_C + p_N = 1$, is satisfied, three different errors of the point estimates of prevalence derived from the EMD and KDE methods, $e_{EMD,1}$, $e_{EMD,2}$, e_{KDE} could be further analysed. The $e_{EMD,1}$ error captures the deviation of the mixture from the convex combination of the two reference distributions. The $e_{EMD,2}$ error is the EMD between the mixture and a model cumulative density function (CDF) based on the two independently estimated not normalised prevalence values p_C^{EMD} and p_N^{EMD}. The e_{KDE} error is the sum of squared residuals (multiplied by the Gaussian kernels bandwidth) from the least-square fitting procedure, which forms part of the KDE method. In order to interpret the values of the errors, we compared them with 100,000 bootstrapped error values (as in all other computations we use 100 mixtures * 1,000 bootstraps). The bootstrap samples are generated using the two reference populations and their composition is based on the initial point estimate of prevalence. In this way, we compare the error value of the investigated mixture sample with 100,000 values from a model that *explicitly* assumes there are only two reference populations (i.e. $p_C + p_N = 1$). This approach allows us to check how likely is the observed error value to occur in the model for a given sample size and reference populations. The obtained bootstrap p-values are the number of bootstrapped modelled errors that are higher than the sample error and can be interpreted as the probability that the observed values of $e_{EMD,1}$, $e_{EMD,2}$ and e_{KDE} are a result of the sampling error. The bootstrap p-values are equivalent to p-values of a traditional statistical test [8,22].

Supplementary Figure 5 shows an example of a mixture of two populations ($p_C + p_N = 1$) and an example of a mixture of three populations (with the third population constituting 7.5% of the mixture, $p_C + p_N + 0.075 = 1$). It illustrates how the $e_{EMD,1}$, $e_{EMD,2}$ and e_{KDE} errors could be used to test the assumption that the mixture is composed of only two populations. In the first case where the mixture is composed of just two populations, the observed $e_{EMD,1}$, $e_{EMD,2}$ and e_{KDE} values are small, and when compared with the 100,000 bootstrapped error values, they indicated that there is high chance: p=0.19 ($e_{EMD,1}$), p=0.95 ($e_{EMD,2}$) and p=0.96 (e_{KDE}) of observing them due to the sampling error in the mixture sample. In the second case where the mixture is composed of three populations, comparison of the observed $e_{EMD,1}$, $e_{EMD,2}$ and e_{KDE} values with the bootstrapped values shows that they are unlikely to be a result of the sampling error: p=0.0001 ($e_{EMD,1}$), p<1e-5 ($e_{EMD,2}$) and p=0.003 (e_{KDE}). In fact, the value of $e_{EMD,2}$ is smaller than any of the bootstrapped error values. However, the figure shows only one particular example and the performance of the methods will depend on the mixture composition (contribution of the other populations) and features of the reference and the other populations.



Supplementary Figure 5: Worked examples of checking the mixture assumption, $p_C + p_N = 1$. A: An example of a mixture that consists of two populations and an example of a mixture that consists of three populations. Left: mixture of two reference populations ($p_C^* = 0.5$, $p_N^* = 0.5$). Right: mixture where the third population has a small contribution ($p_C^* = 0.5$, $p_N^* = 0.425$, $p_3^* = 0.075$; R_3 is a truncated normal distribution with mean 0.17 and std 0.025). B: Illustration of the methods: $e_{EMD,1}$, deviation from collinearity between the two reference distributions and the mixture; $e_{EMD,2}$, EMD between the mixture and a model CDF based on the two independently estimated prevalence values; e_{KDE}, the sum of squared residuals of the final fit.