Bayesian clustering with uncertain data

Kath Nicholls[1,2], Paul D W Kirk[1,2,3], and Chris Wallace[1,2]

[1]Cambridge Institute of Therapeutic Immunology and Infectious Disease
[2]MRC Biostatistics Unit, University of Cambridge
[3]Cancer Research UK Cambridge Centre, Ovarian Cancer Programme, University of Cambridge

Abstract

Clustering is widely used in bioinformatics and many other fields, with applications from exploratory analysis to prediction. Many types of data have associated uncertainty or measurement error, but this is rarely used to inform the clustering.

We present Dirichlet Process Mixtures with Uncertainty (DPMUnc), an extension of a Bayesian nonparametric clustering algorithm which makes use of the uncertainty associated with data points. We show that DPMUnc out-performs existing methods on simulated data. We cluster immune-mediated diseases (IMD) using GWAS summary statistics, which have uncertainty linked with the sample size of the study. DPMUnc separates autoimmune from autoinflammatory diseases and isolates other subgroups such as adult-onset arthritis.

We additionally consider how DPMUnc can be used to cluster gene expression datasets that have been summarised using gene signatures. We first introduce a novel procedure for generating a summary of a gene signature on a dataset different to the one where it was discovered. Since the genes in the gene signature are unlikely to be as strongly correlated as in the original dataset, it is important to quantify the variance of the gene signature for each individual. We summarise three public gene expression datasets containing patients with a range of IMD, using three relevant gene signatures. We find association between disease and the clusters returned by DPMUnc, with clustering structure replicated across the datasets.

The significance of this work is two-fold. Firstly, we demonstrate that when data has associated uncertainty, this uncertainty should be used to inform clustering and we present a method which does this, DPMUnc. Secondly, we present a procedure for using gene signatures in datasets other than where they were originally defined. We show the value of this procedure by summarising gene expression data from patients with immune-mediated diseases using relevant gene signatures, and clustering these patients using DPMUnc.

Author Summary

Identifying groups of items that are similar to each other, a process called clustering, has a range of applications. For example, if patients split into two distinct groups this suggests that a disease may have subtypes which should be treated differently. Real data often has measurement error

associated with it, but this error is frequently discarded by clustering methods. We propose a clustering method which makes use of the measurement error and use it to cluster diseases linked to the immune system.

Gene expression datasets measure the activity level of all ∼20,000 genes in the human genome. We propose a procedure for summarising gene expression data using gene signatures, lists of genes produced by highly focused studies. For example, a study might list the genes which increase activity after exposure to a particular virus. The genes in the gene signature may not be as tightly correlated in a new dataset, and so our procedure measures the strength of the gene signature in the new dataset, effectively defining measurement error for the summary. We summarise gene expression datasets related to the immune system using relevant gene signatures and find that our method groups patients with the same disease.

Introduction

Grouping items by similarity has a variety of applications in bioinformatics. For example, clustering samples according to any molecular measures may help to identify subgroups of disease[1,2]. In gene expression data, clustering the genes may be useful for inferring gene function using guilt by association and inferring regulatory relationships[3]. Clustering can also be used to cluster diseases themselves. For example, immune-mediated diseases (IMD) have been shown to have a shared genetic basis such that IMD can be clustered according to those shared patterns[4,5].

Commonly used methods, such as k-means[6,7] and mclust[8], assume that the observations being clustered are observed without error, or that any errors are identically distributed. In reality we often have not just data points themselves but also a measure of uncertainty about each observation which has the potential to improve clustering accuracy if exploited. Within the clustering approaches, the Bayesian methods offer some advantages, in their ability to convey uncertainty about the cluster centers and variances as well as the overall clustering, and to simultaneously infer the number of clusters, K. Here, we propose Dirichlet Process Mixtures with Uncertainty (DPMUnc) which adapts Bayesian clustering to the setting where the data points have associated uncertainty.

We show that accounting for uncertainty can alter the clustering solution returned and demonstrate its performance on a range of simulated datasets and real applications. First, we consider clustering IMD using GWAS summary statistics, a problem where the uncertainty associated with each observation generally varies systematically with study sample size. Second, we cluster patients with IMD using gene expression data which can help identify subtypes of disease or predict response to treatment[2]. However, clustering on all ∼20,000 genes can be computationally expensive, and may group patients by disease-irrelevant structure, such as sex or age. Investigations of gene expression in relation to disease or biological processes sometimes produce a gene signature, a list of genes with a shared pattern of gene expression which is relevant to the disease or process. We propose a summary measure of a gene signature, which captures average signature expression and its variance, and show that DPMUnc allows patients to be clustered according to one or more signatures.

Results

DPMUnc model

Our approach follows previous work[9], extended to include both the observations $x_i,\ i = 1 \dots n$ and their respective uncertainty estimates σ_i^2. We model these x_i as noisy observations of some latent variables z_i: $x_i \sim \mathcal{N}\left(z_i, \sigma_i^2\right)$. The latent variables themselves are modelled as coming from a Dirichlet Process Gaussian Mixture Model (DPGMM). We follow the approach of[10], which takes the infinite limit of a Gaussian Mixture Model and can be shown to be equivalent to using Dirichlet Processes directly[10,11].

In a finite Gaussian Mixture Model the points are generated from K clusters, where points within each cluster are i.i.d. Gaussian variables with some mean μ_k and variance ρ_k^2:

$$z_i | c_i = k \sim \mathcal{N}(\mu_k, \rho_k^2)$$

where the indicator variables c_i show which cluster a variable belongs to. We give these indicator variables a standard categorical distribution with the weights for each category π_k having a Dirichlet prior:

$$p(c_i = k) = \pi_k; \quad \pi_1, \dots, \pi_K \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K)$$

where $\sum_{k=1}^K \pi_k = 1$.

Finally, we define priors

$$\alpha \sim \text{Gamma}(a_0, b_0)$$
$$\mu_k, \rho_k^2 \sim \text{Normal-Inverse Gamma}(\mu_0, \kappa_0, \alpha_0, \beta_0)$$

where $a_0 = 3, b_0 = 4$, μ_0 is the empirical mean of the observed data and $\alpha_0, \beta_0, \kappa_0$ are hyperparameters that should be adjusted according to the spread of the data and the expected cluster structure. Some further discussion of how these values may be chosen is given in the Supplementary Information.

The joint density factorises as follows:

$$p(\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\rho}, \alpha) = p(\boldsymbol{\phi}) p(\boldsymbol{c}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha) p(\alpha) \prod_i p(x_i|\sigma_i^2, z_i) p(z_i|\phi_{c_i}) \tag{1}$$

as represented in Figure 1. We now let K tend to infinity, yielding a Gaussian Mixture Model with no constraint on the number of clusters which is equivalent to a Dirichlet Process Gaussian Mixture Model[10,11].

We use a Gibbs sampling algorithm to sample from the posterior $p(\boldsymbol{c}|\boldsymbol{x}, \boldsymbol{\sigma^2})$ according to the conditional probability distributions as set out in Supplementary Information and as described in Supplementary Figure S9.

To more directly compare the effect of adjusting for the uncertainty of the points, we also run DPMUnc with adjusted versions of the datasets where the observations themselves were

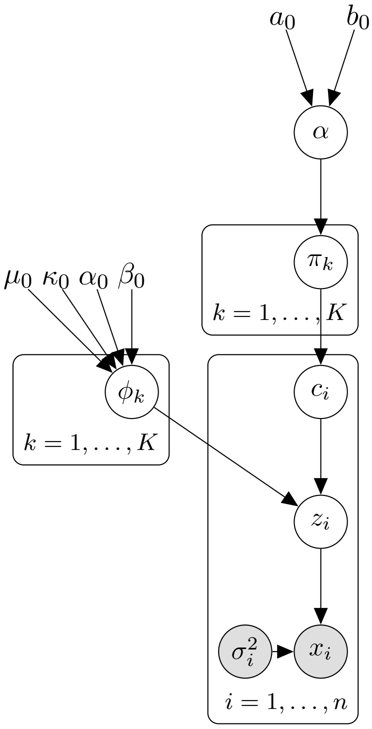

Fig. 1: Plate diagram of Gaussian Mixture Model. When $K \to \infty$ this is the Dirichlet Process Gaussian Mixture Model that is the basis of our model. In the plate diagram, each rectangle (plate) corresponds to a group of variables, and the text at the bottom of each plate shows how many copies of that plate are required for the full model. Circles indicate random variables, which are shaded if the random variable is observed. The variables without circles are hyperparameters for the model.

untouched, but the uncertainty estimates were shrunk to 0. Thus the rest of the DPMUnc works as normal, but the latent variables are essentially fixed to be equal to the observed data points throughout and so the only difference compared to DPMUnc is that no adjustment for uncertainty takes place. We call this version of the method DPMZeroUnc.

Simulated study

We first used simulated data to explore the effects of differing levels of uncertainty associated with each observation and noise around each cluster location on DPMUnc. Detailed discussion of two illustrative examples is presented in the Supplementary Information. Briefly, we found that including estimates of uncertainty tended to produce solutions with smaller cluster variances, and with cluster means corresponding to a weighted mean of observations, in which observations with lower uncertainty were given greater weight, rather than unweighted means when uncertainty was ignored.

Using the larger simulations, DPMUnc outperformed its comparators kmeans and mclust by a small margin when observation noise and cluster variance were small, which increased with increasing cluster variance or observation noise (Figure 2). The methods differed most in their ability to infer K. When the true $K = 3$, DPMUnc and DPMZeroUnc tended to select $K = 3$ or $K = 4$, whilst kmeans and, to a greater extent, mclust, tended to merge clusters and underestimate K, often selecting $K = 1$ as cluster noise increased (Supplementary Figure S7).

As a Bayesian method, DPMUnc not only provides a clustering of the data, but can quantify the uncertainty about the clustering. This can be captured using the posterior similarity matrix (PSM) whose entry $\text{PSM}_{i,j}$ is the proportion of samples from the MCMC sampling in which observation i and observation j were placed in the same cluster[12].

We assessed how well-calibrated the posterior similarity scores were, by binning PSM values into deciles, and examining the proportion of trait pairs in each bin that were truly in the same cluster. We found the two were generally similar, suggesting that DPMUnc was relatively well-calibrated (Supplementary Figure S8).

Clustering immune-mediated diseases using GWAS summary statistics

Methodology has been developed to summarise genetic associations shared by multiple diseases using multi-dimensional polygenic scores, such as the summary of IMD by 13 such scores[13]. The paper also calculated these 13-dimensional scores for 1,230 independent GWAS datasets. We ran DPMUnc on 45 traits from these published data that showed significant differences to controls according to the original study (Figure 3). GWAS summary statistics include both an estimate of the strength of association and the uncertainty associated with that estimate. The uncertainty is propagated through to the scores, which is influenced heavily by sample size so each trait has similar uncertainty across all the scores, while uncertainty varies between traits (Figure 3).

Six clusters were identified, summarised as classic autoimmune diseases and most arthitides (A), a null cluster with observations lying close to zero or with sufficiently high uncertainty that they might truly lie near zero (B), classic autoinflammatory diseases and coeliac disease (C), multiple sclerosis (D), adult onset arthritis (E), and eosophilic granulomatisos granulomatosis with polyangiitis (EGPA) (F).

The posterior similarity matrix in Figure S10 shows relatively high confidence in the clusters. In particular, even with the uncertainty associated with the location of the rare disease EGPA, the two EGPA subtypes form their own cluster in a high percentage of the samples, perhaps reflecting their extreme location on PC13 which was shown to be related to eosinophils[13]. In contrast, whilst coeliac disease is clustered with other inflammatory bowel conditions (Crohn's disease and Ulcerative Colitis), it is much less confidently placed in this cluster than the other traits, perhaps reflecting that it is generally considered an autoimmune disease.

Multiple sclerosis (MS) is the only disease to cluster alone, which perhaps reflects that MS is unusual in having both autoinflammatory and autoimmune features[14], although the PSM does show both neuromyelitis optica (NMO) IgG^- and systemic JIA have some non-zero probability of clustering with MS.

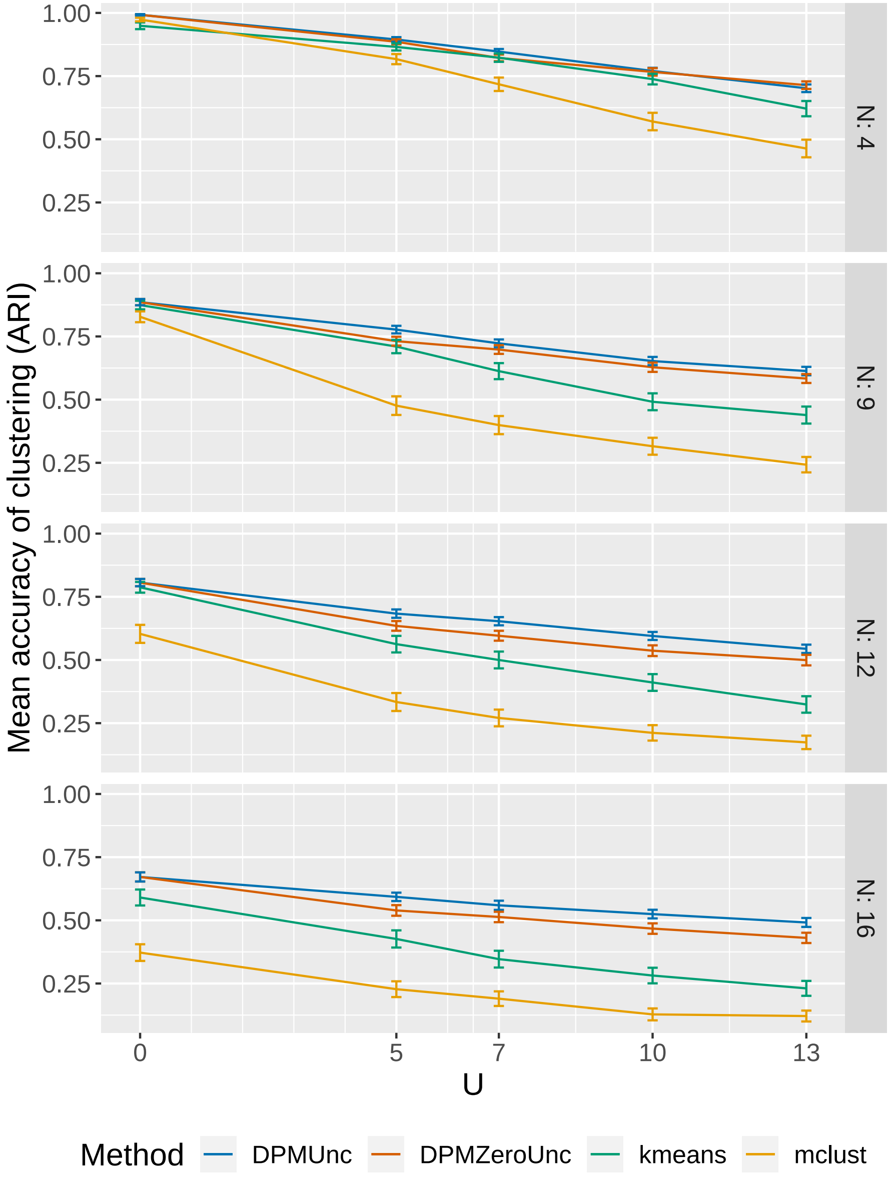

Fig. 2: Accuracy on simulated datasets. The first row of plots has the lowest noise N around the cluster mean, and the bottom row has the highest noise. Increasing uncertainty, corresponding to greater difference between latent data and the observed data is shown on the x-axis. The accuracy of the clustering is given by the Adjusted Rand Index (ARI) between the true clustering and the inferred clustering. Higher values of ARI are better.

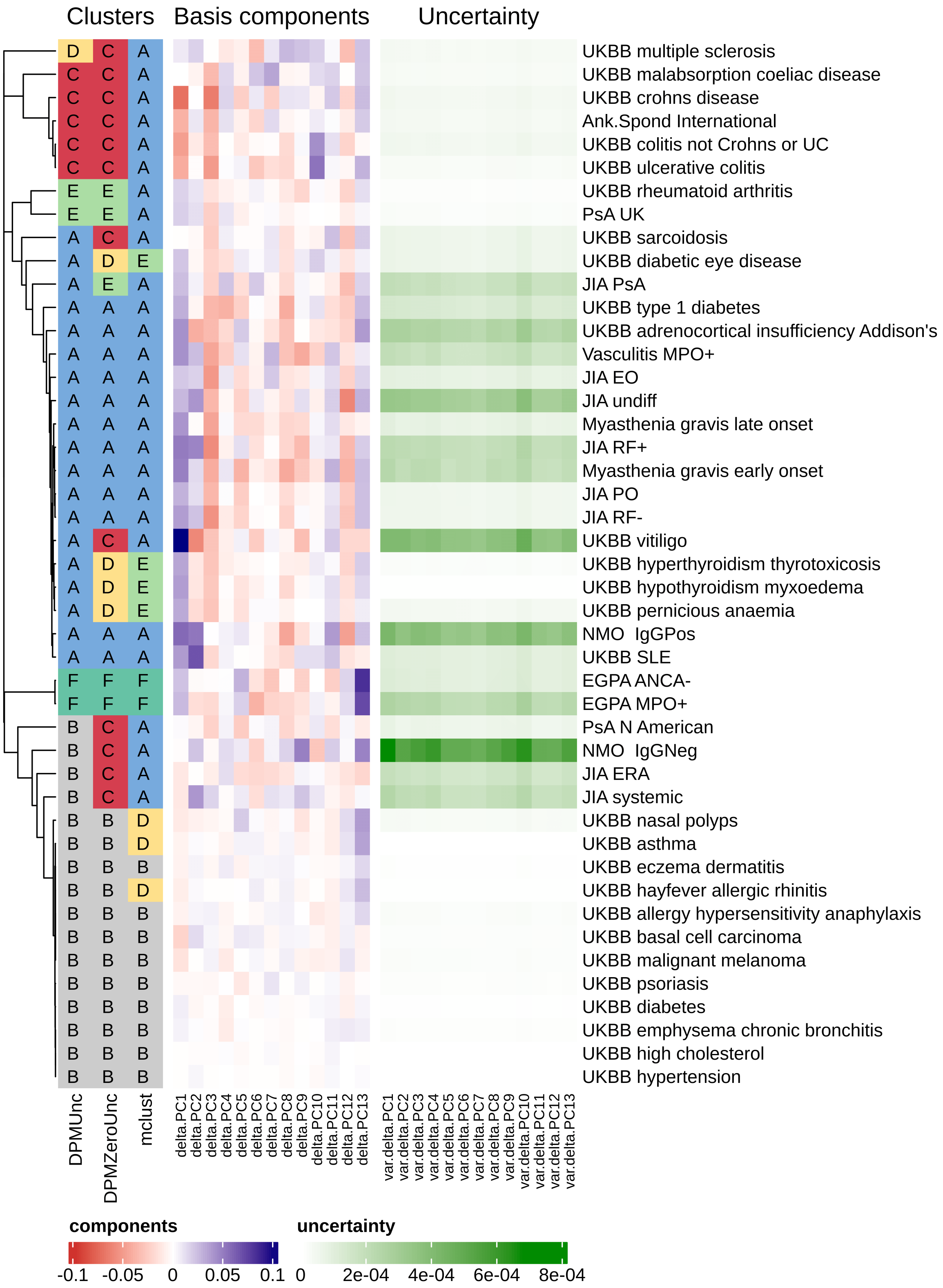

Fig. 3: Clusters inferred by DPMUnc, DPMZeroUnc and mclust alongside the polygenic GWAS scores and the associated uncertainty. The clusters of DPMZeroUnc and mclust have been coloured to best match the colours of DPMUnc. The dendrogram on the left is formed by applying hierarchical clustering on the posterior similarity matrix from DPMUnc, using complete-linkage.

	Ferreira[16]	Lyons[20]	Chaussabel[19]
Healthy controls (HC)	88	25	12
Systemic lupus erythematosus (SLE)	25	13	103
Juvenile idiopathic arthritis (JIA)			47
Melanoma (MEL)			39
Transplant (RET)			37
Type 1 diabetes (T1D)	64		20
Vasculitis		30	
Microscopic Polyangiitis (VWP)		*8*	
Wegener's Granulomatosis (VMG)		*22*	
Infection			40
Infection (E. Coli) (EColi)			*22*
Infection (S. Aureus) (SAureus)			*18*
Total	177	68	298

Table 1: Sample counts in each gene expression dataset. Counts for a subtype of a disease are in italics.

Clustering patients by gene expression signatures

Gene expression signatures have been linked to several immune-mediated diseases, and we identified three such signatures from recent literature:

$\boldsymbol{Sig}_{\mathbf{NK}}$ contains genes enriched for expression in NK cells and has been associated with T1D[15]

$\boldsymbol{Sig}_{\mathbf{IFN}}$ contains genes in the type I interferon pathway, commonly found to show increased expression in SLE patients[16]

$\boldsymbol{Sig}_{\mathbf{CD8}}$ genes associated with CD8 T cell exhaustion, which have been shown to predict prognosis in inflammatory bowel disease[17,18]

We attempted to cluster patients with a range of IMD in three gene expression datasets[16,19,20] (Table 1). For each individual, we summarised expression levels of each signature with a mean level and its associated uncertainty, as described in Methods. For some signatures, we anticipated association with specific diseases (such as the interferon signature and SLE) whilst for others we had less clear prior hypotheses. Whilst we did not expect every signature or every clustering to associate with disease, because any clusters found might reflect disease activity, sex or ethnicity, we considered that clusterings which discriminated between diseases would be meaningful, and so assessed the association of each clustering with disease using Fisher's exact test.

We first used DPMUnc to cluster datasets by individual signatures, and found clusters in each dataset which showed association with disease in the majority of cases (7/9). All datasets identified a "high" cluster for Sig_{IFN} which contained a majority of SLE cases. Clustering by Sig_{CD8} also found a "high" cluster which contained a majority of T1D cases in both datasets

(Ferreira and Lyons). For Sig_{NK}, the Ferreira dataset identified a low cluster containing primarily T1D cases, as expected given the published associated of this signature with T1D[15], but whilst evidence for differential clustering by disease was found in Chaussabel, there was no enrichment of T1D in the "low" cluster. This difference may reflect time since diagnosis - T1D cases in Ferreira were predominantly newly diagnosed, while the samples in the original study that identified the signature were from children before and around the time of seroconversion. Only limited details are given about the Chaussabel T1D samples[19], but since by definition T1D cases are only newly diagnosed for a short period of time, it is likely these were longer standing cases. Thus this signature may represent an active disease process that resolves in long standing T1D cases.

We then clustered the Ferreira and Chaussabel datasets, which had shown association with each signature, across all three signatures simultaneously. We found that in the case of Ferreira, the multiple-signature clustering discriminated more powerfully between diseases than any individual signature clustering (Figure 5). However, in the case of Chaussabel, while the SLE group remained identifiable with high Sig_{IFN} values, the T1D cluster that had been identified in Sig_{CD8} clustering alone, was folded into the main cluster with the healthy controls.

Discussion

Taking uncertainty into account can have an impact on clustering. We illustrated with simulated data how the cluster mean inferred by DPMUnc is closer to the points that have lower uncertainty, whereas methods like mclust and k-means place the cluster mean close to the empirical mean of all the points, and also how DPMUnc shifts the latent data points towards the inferred cluster mean, which results in a smaller cluster variance. In a more complicated dataset this could have knock-on consequences, perhaps excluding more distant points from joining the cluster, or splitting one cluster into two. On a range of simulated datasets, DPMUnc out-performed existing methods and the posterior similarity values were shown to be relatively well calibrated, so that if two points have posterior similarity of p, this roughly translates to probability p that they are in the same cluster in the true clustering.

Our clustering of diseases by the engineered GWAS features recapitulated known divisions and relationships in the immune-mediated diseases. It also suggests further uses for clustering with uncertainty, such as clustering individuals using sets of polygenic risk scores, where the uncertainty in the PRS coefficients could be exploited rather than ignored.

Gene signatures may be defined on one dataset with ideal experimental conditions, such as a timecourse study of the effect of interferon-β. To then use them on a new dataset with more complex structure, such as a dataset with patients with a mix of immune-mediated diseases, typically requires some form of dimension reduction such as WGCNA or PCA used on the new dataset, either using just those genes from the signature[16] or using all genes[18] with the hope that one WGCNA module or variable in PCA will coincide with the signature. Our proposed method of summarising a gene expression across a signature extends the possible uses of signatures, and since the signature may be weaker on a dataset with more complex structure, it is crucial to take

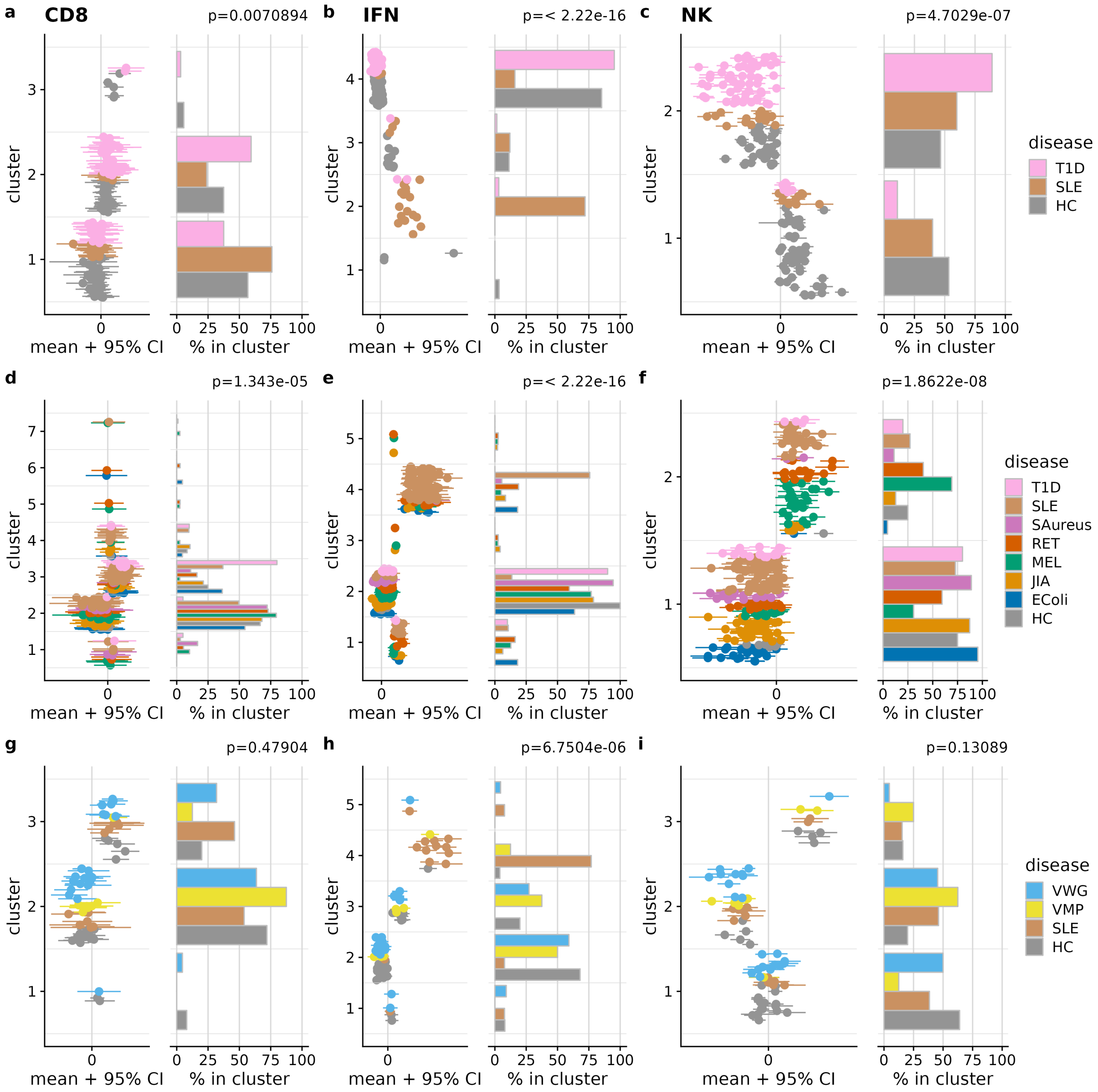

Fig. 4: Clustering of samples in 3 gene expression datasets (rows) according to 3 gene signatures (columns). (a-c) Ferreira (d-f) Chaussabel (g-i) Lyons. In each panel, the left plot shows the observed data, and the right plot shows the fraction of individuals assigned to each cluster. The p value shown relates to the null hypothesis that cluster membership is independent of disease.

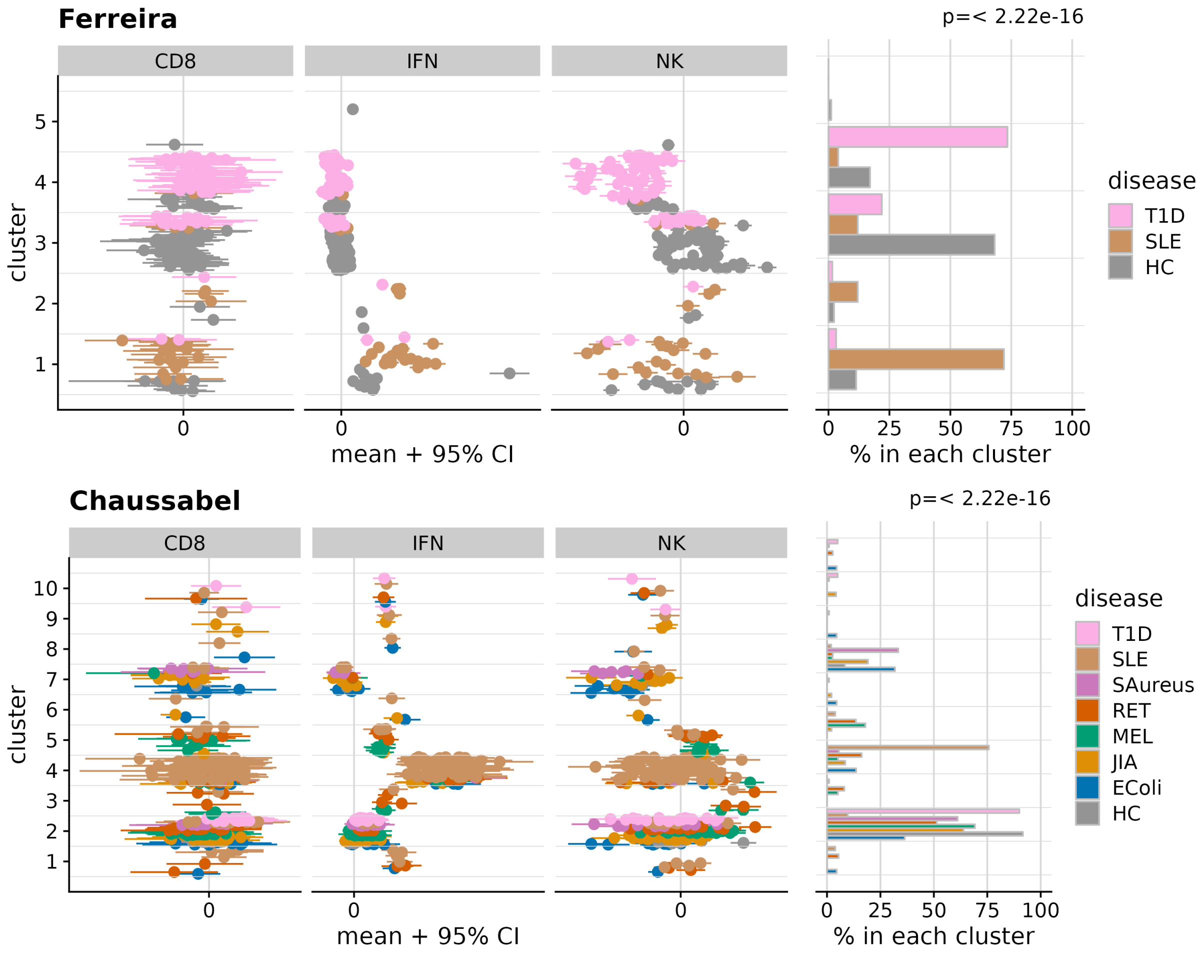

Fig. 5: Clustering of samples in 2 gene expression datasets (rows) using all 3 gene signatures. In each panel, the left plot shows the observed data, and the right plot shows the fraction of individuals assigned to each cluster. The p value shown relates to the null hypothesis that cluster membership is independent of disease.

the variability of the signature into account, which DPMUnc allows. In particular, it allows us to cluster patients simultaneously according to multiple signatures. In the case of the Ferreira dataset, this led to a clearer separation of the two disease and control groups, emphasizing the potential utility of considering multiple signatures, which provide multiple views of the same data.

Methods

Inference from DPMUnc

Simulations

We simulated a variety of datasets with different characteristics to explore the behaviour of DPMUnc in comparison to other methods when the "truth" was known.

The simulated datasets each consist of thirty points spread across three clusters, and broadly follow the model used by DPMUnc outlined above. The simulation process includes two parameters which are varied between simulations: U controls the magnitude of observation uncertainty and N controls the level of noise around the cluster means, i.e. the levels of observational and biological variation respectively. The process for simulating data is:

1. Cluster means are the points $\mu_1 = (10, 0)$, $\mu_2 = (0, 10)$, $\mu_3 = (\alpha, \alpha)$ where $\alpha = 10\frac{1+\sqrt{3}}{2} \approx 14$ was chosen so that the cluster means are equidistant.

2. Points were independently and uniformly assigned to clusters i.e. $c_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}\{1, 2, 3\}$ for $i = 1, \dots, 30$.

3. Latent observations were simulated around the cluster mean, with variance N i.e. $z_i \mid c_i = k \sim \mathcal{N}(\mu_k, N)$

4. Uncertainty for the points was simulated using $\sigma_i^2 \overset{\text{i.i.d.}}{\sim} \Gamma(k = 1/2, \theta = 2U)$.

5. Observed data points were simulated about the latent points: $x_i \sim \mathcal{N}\left(z_i, \sigma_i^2\right)$

We simulated for 100 different seeds each combination of $U = 5, 7, 10, 13$ and $N = 4, 9, 12, 16$. A grid of simulated datasets is shown in Figure S6 showing the range of the datasets.

We ran DPMUnc and DPMZeroUnc with 10 thousand iterations, saving only every 10th iteration to avoid autocorrelation and discarding the first half of the iterations as burn-in.

Prior for cluster parameters

We used the prior hyperparameters $\alpha_0 = 2, \kappa_0 = 0.5$. μ_0 is set to be the empirical mean of all the data. β_0 is set to be 0.2 multiplied by the variance of the variable, thus scaling β_0 by the variance of the variable. In datasets with multiple variables, we use the mean of the variances to scale β_0. If a dataset had variables on vastly different scales, this would not be an appropriate choice of prior but all the datasets in this paper have variables on comparable scales.

Comparing to other methods

We used simulated data to examine the performance of DPMUnc. Simulated data provides access to the true latent observations and so we could use the clustering ability on the latent data as a guide to the best performance possible by DPMUnc: no matter how good the adjustment made for uncertainty of the points is, DPMUnc cannot be expected to do better than its performance on the true latent data. We also used two widely used clustering methods as benchmarks, using the adjusted rand index (ARI)[21] to quantify the accuracy.

mclust[8] is a Gaussian Mixture Model (GMM) which does not account for uncertainty and which requires K to be specified. We chose K by optimising the Bayesian Information Criterion.

k-means[6,7] can be shown to be equivalent to a GMM with spherical clusters. The R package *cluster*[22] provides a function *clusGap* which calculates a goodness of clustering measure called the gap statistic[23], which measures within-cluster similarity. We use this measure to choose the optimal value of k, the number of clusters.

We also ran DPMUnc with adjusted versions of the datasets where the uncertainty estimates were shrunk to 0. Thus the latent variables are essentially fixed to be equal to the observed data points throughout. We call this version of the method DPMZeroUnc.

GWAS summary statistics features

We downloaded Table S10 from[13] and kept the 186 traits which differed significantly from a control dataset (FDR$<$ 0.01), and focused on disease traits. To avoid near duplicates, we retained only one set of results from UK Biobank (UKBB) by excluding the GeneAtlas analysis, and retaining the Neale analyses which had a greater representation of rarer immune traits, leaving us with 45 traits in our final dataset.

We ran DPMUnc and DPMZeroUnc with 100 million iterations, saving only every 1000th iteration to avoid autocorrelation and discarding the first half of the iterations as burn-in.

Gene expression signatures

The genes representing each signature were identified as follows

Sig

$\boldsymbol{Sig}_{\textbf{NK}}$ consists of 87 genes (Supplementary File S6 from[15]).

$\boldsymbol{Sig}_{\textbf{IFN}}$ consists of 56 genes identified through differential expression analysis using interferon stimulation of PBMCs and marked as discriminatory in SLE according to Supplementary Table 2[16].

$\boldsymbol{Sig}_{\textbf{CD8}}$ consists of 12 genes from module 8 and module 9 in[17] Supplemental Table 1 which showed reduced expression with log fold change < -3 only in exhausted CD8 T cells. The signature was defined in mice and we used BioMart to convert from murine (GRCm39) genes to human (GRCh38.p13) genes[24,25].

The selected signature genes are listed in full in Table S1.

MCMC summarisation approach

Whilst DPMUnc and DPMZeroUnc provide samples from a posterior distribution, it is useful to find a single summary clustering for downstream analyses. We first calculate the posterior similarity matrix (PSM). Each entry $\mathrm{PSM}_{i,j}$ is the proportion of posterior samples in which observation i and observation j were placed in the same cluster. Secondly, we calculate the clustering of the PSM that optimises the posterior expected adjusted rand index (PEAR) with the true clustering (maxpear[12]), i.e. the clustering that seems the most compatible with the true clustering. This requires a search strategy to explore the space of possible clusters, and we use all possible clusters found via complete hierarchical clustering, using $1 - \mathrm{PSM}$ as a distance matrix, using the mcclust and mcclust.ext packages in R.

Preprocessing gene expression data

We first used a variance stabilising transformation[26] *vsn2* to minimise the relationship between average gene expression and the variance of gene expression (Figure S4). We then used median absolute deviation (MAD) scale normalization to make the distribution of expression values similar between genes, dividing all values for gene g by the MAD of the control samples for gene g. For each sample, we then summarised using the mean and variance of the normalised expression value, with the summary measures taken across all genes in a signature.

We ran DPMUnc and DPMZeroUnc with 20 million iterations, saving only every 500th iteration to avoid autocorrelation and discarding the first half of the iterations as burn-in.

References

1. Jiang D, Tang C, Zhang A. Cluster Analysis for Gene Expression Data: A Survey. IEEE Transactions on Knowledge and Data Engineering. 2004 Nov;16(11):1370–1386.

2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science. 1999 Oct;286(5439):531–537.

3. Saelens W, Cannoodt R, Saeys Y. A Comprehensive Evaluation of Module Detection Methods for Gene Expression Data. Nature Communications. 2018 Mar;9(1):1–12.

4. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive Sharing of Genetic Effects in Autoimmune Disease. PLoS genetics. 2011 Aug;7(8):e1002254.

5. Majumdar A, Haldar T, Bhattacharya S, Witte JS. An Efficient Bayesian Meta-Analysis Approach for Studying Cross-Phenotype Genetic Associations. PLoS genetics. 2018 Feb;14(2):e1007139.

6. Lloyd S. Least Squares Quantization in PCM. IEEE Transactions on Information Theory. 1982 Mar;28(2):129–137.

7. MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations. In: Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics. Univ. California Press, Berkeley, Calif.; 1967. p. 281–297.

8. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. The R Journal. 2016;8(1):289–317.

9. Escobar MD, West M. Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association. 1995 Jun;90(430):577–588.

10. Rasmussen CE. The Infinite Gaussian Mixture Model. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. NIPS'99. Cambridge, MA, USA: MIT Press; 1999. p. 554–560.

11. Neal RM. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics. 2000 Jun;9(2):249.

12. Fritsch A, Ickstadt K. Improved Criteria for Clustering Based on the Posterior Similarity Matrix. Bayesian Analysis. 2009 Jun;4(2):367–391.

13. Burren OS, Reales G, Wong L, Bowes J, Lee JC, Barton A, et al. Genetic Feature Engineering Enables Characterisation of Shared Risk Factors in Immune-Mediated Diseases. Genome Medicine. 2020 Nov;12(1):106.

14. Wootla B, Eriguchi M, Rodriguez M. Is Multiple Sclerosis an Autoimmune Disease? Autoimmune Diseases. 2012;2012:1–12.

15. Xhonneux LP, Knight O, Lernmark Å, Bonifacio E, Hagopian WA, Rewers MJ, et al. Transcriptional Networks in At-Risk Individuals Identify Signatures of Type 1 Diabetes Progression. Science translational medicine. 2021 Mar;13(587):eabd5666.

16. Ferreira RC, Guo H, Coulson RMR, Smyth DJ, Pekalski ML, Burren OS, et al. A Type I Interferon Transcriptional Signature Precedes Autoimmunity in Children Genetically at Risk for Type 1 Diabetes. Diabetes. 2014 Jul;63(7):2538–2550.

17. Wherry EJ, Ha SJ, Kaech SM, Haining WN, Sarkar S, Kalia V, et al. Molecular Signature of CD8+ T Cell Exhaustion during Chronic Viral Infection. Immunity. 2007 Oct;27(4):670–684.

18. McKinney EF, Lee JC, Jayne DRW, Lyons PA, Smith KGC. T-Cell Exhaustion, Co-Stimulation and Clinical Outcome in Autoimmunity and Infection. Nature. 2015 Jul;523(7562):612–616.

19. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. Immunity. 2008 Jul;29(1):150–164.

20. Lyons PA, McKinney EF, Rayner TF, Hatton A, Woffendin HB, Koukoulaki M, et al. Novel Expression Signatures Identified by Transcriptional Analysis of Separated Leucocyte Subsets in Systemic Lupus Erythematosus and Vasculitis. Annals of the Rheumatic Diseases. 2010 Jun;69(6):1208–1213.

21. Hubert L, Arabie P. Comparing Partitions. Journal of Classification. 1985 Dec;2(1):193–218.

22. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions; 2022. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source). Available from: https://CRAN.R-project.org/package=cluster.

23. Tibshirani R, Walther G, Hastie T. Estimating the Number of Clusters in a Data Set via the Gap Statistic. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2001;63(2):411–423.

24. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: A Powerful Link between Biological Databases and Microarray Data Analysis. Bioinformatics. 2005 Aug;21(16):3439–3440.

25. Durinck S, Spellman PT, Birney E, Huber W. Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package biomaRt. Nature Protocols. 2009;4(8):1184–1191.

26. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. Bioinformatics. 2002 Jul;18(suppl_1):S96–S104.

Data Availability Statement The datasets were derived from sources in the public domain, as follows:

- Ferreira from https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1724
- Lyons from https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-145
- Chaussabel from https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-11907

Software

We make the implementation of DPMUnc, and the code to generate the analyses in this manuscript, available in github repositories:

- DPMUnc software https://github.com/nichollskc/DPMUnc
- DPMUnc simulation study https://github.com/nichollskc/DPMUnc_simulations
- GWAS dataset analysis https://github.com/nichollskc/DPMUnc_GWAS
- Gene expression dataset analysis https://github.com/nichollskc/clusteringPublicGeneExpr

Acknowledgements

This work is supported by the Wellcome Trust (WT220788, WT220024) and the Medical Research Council (MC UU 00002/4) and supported by the NIHR Cambridge BRC (BRC-1215-20014). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

Conflicts of interest

CW receives research funding from GSK and MSD for an unrelated project and is a part-time employee of GSK. These companies had no input into this study.

Author contributions

KN: Data curation, investigation, visualization, writing - original draft
PK: methodology, writing - review and editing
CW: supervision, investigation, visualization, writing - original draft