# Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions

Yang Nan [a,*], Javier Del Ser [d,e], Simon Walsh [a], Carola Schönlieb [f], Michael Roberts [f,g], Ian Selby [h], Kit Howard [i], John Owen [i], Jon Neville [i], Julien Guiot [j,k], Benoit Ernst [j,k], Ana Pastor [l], Angel Alberich-Bayarri [l], Marion I. Menzel [m,n], Sean Walsh [o], Wim Vos [o], Nina Flerin [o], Jean-Paul Charbonnier [p], Eva van Rikxoort [p], Avishek Chatterjee [q], Henry Woodruff [q], Philippe Lambin [q], Leonor Cerdá-Alberich [r], Luis Martí-Bonmatí [r], Francisco Herrera [s,t,#], Guang Yang [a,b,c,#,*]

[a] National Heart and Lung Institute, Imperial College London, London, Northern Ireland UK
[b] Cardiovascular Research Centre, Royal Brompton Hospital, London, Northern Ireland UK
[c] School of Biomedical Engineering & Imaging Sciences, King's College London, London, Northern Ireland UK
[d] Department of Communications Engineering, University of the Basque Country UPV/EHU, Bilbao 48013, Spain
[e] TECNALIA, Basque Research and Technology Alliance (BRTA), Derio 48160, Spain
[f] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, Northern Ireland UK
[g] Oncology R&D, AstraZeneca, Cambridge, Northern Ireland UK
[h] Department of Radiology, University of Cambridge, Cambridge, Northern Ireland UK
[i] Clinical Data Interchange Standards Consortium, Austin, TX, United States of America
[j] University Hospital of Liège (CHU Liège), Respiratory medicine department, Liège, Belgium
[k] University of Liege, Department of clinical sciences, Pneumology-Allergology, Liège, Belgium
[l] QUIBIM, Valencia, Spain
[m] Technische Hochschule Ingolstadt, Ingolstadt, Germany
[n] GE Healthcare GmbH, Munich, Germany
[o] Radiomics (Oncoradiomics SA), Liège, Belgium
[p] Thirona, Nijmegen, The Netherlands
[q] Department of Precision Medicine, Maastricht University, Maastricht, The Netherlands
[r] Medical Imaging Department, Hospital Universitari i Politècnic La Fe, Valencia, Spain
[s] Department of Computer Sciences and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI) University of Granada, Granada, Spain
[t] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

## ARTICLE INFO

## ABSTRACT

Removing the bias and variance of multicentre data has always been a challenge in large scale digital healthcare studies, which requires the ability to integrate clinical features extracted from data acquired by different scanners and protocols to improve stability and robustness. Previous studies have described various computational approaches to fuse single modality multicentre datasets. However, these surveys rarely focused on evaluation metrics and lacked a checklist for computational data harmonisation studies. In this systematic review, we summarise the computational data harmonisation approaches for multi-modality data in the digital healthcare field, including harmonisation strategies and evaluation metrics based on different theories. In addition, a comprehensive checklist that summarises common practices for data harmonisation studies is proposed to guide researchers to report their research findings more effectively. Last but not least, flowcharts presenting possible ways for methodology and metric selection are proposed and the limitations of different methods have been surveyed for future research.

* Corresponding authors.
  *E-mail addresses:* y.nan20@imperial.ac.uk (Y. Nan), g.yang@imperial.ac.uk (G. Yang).
# Francisco Herrera and Guang Yang are co-last authors of this work.

## 1. Introduction

Computational biomedical research aims to advance digital healthcare and biomedical studies by developing computational models that improve the precise diagnosis of disease spectrum, analysis of gene expressions or time series data (e.g., electroencephalograms and electrocardiograms). These models are designed to discover novel risk biomarkers, predict disease progression, design optimal treatments, and identify new drug targets for applications such as cancer, pulmonary disease, and neurological disorders. Whilst a well-performed model should have characteristics of high performance, robustness, explainability, and reproducibility, it faces the issue that the bias of distribution between different datasets dramatically increases the difficulty of developing models from large-scale studies. Although data harmonisation is needed with almost any kind of medical data, automated methods have been extensively used for medical images, gene expression analysis, with the rest of the modalities being ignored or harmonised manually. Studies have shown that machine learning based approaches, especially deep neural networks, are highly sensitive to the distribution of training data. Therefore, there is an urgent need to develop approaches that can integrate the device/site-invariant information from multiple datasets. To address this issue, researchers established standard acquisition protocols [1 3] or definitions [4,5] to help data collectors to glean standardised data. For instance, Delbeke et al. [2] recommended an acquisition protocol for F-FDG Positron emission tomography/computerised tomography imaging (PET/CT), and Simon et al. [3] presented a standardised MR imaging protocol for multi-sclerosis. Schmidt et al. [4,5] mainly focused on integrating the data from routine health information systems, including conducting manual harmonisation and rule-based alignment of electronic data. Although these acquisition protocols could effectively reduce the cohort bias (non-biological variances in cross-scanner/site data), they were limited in assisting prospective studies because most studies were retrospective and could not be re-acquired with the same standard. In addition, a non-standardised acquisition protocol is needed for personalised digital healthcare sometimes. Therefore, it is imperative to explore a computational method to harmonise multicentre datasets.

Although some surveys of computational data harmonisation have been released [6,7,10], such as MRI (magnetic resonance imaging) [11] or CT (computerised tomography) harmonisation, these surveys only explored methods of single modality or application and rarely focused on evaluation metrics and research guidance (shown in Table 1). Moreover, there is a lack of a checklist that can summarise the common practice and give guidance for methodology selection and development for computational data harmonisation studies. This survey summarises the computational data harmonisation strategies for multimodal data in the digital healthcare field in terms of methodologies, evaluations, and applications. Our paper covers three main areas (i.e., gene expression, radiomics, and pathology), with over 96 qualified papers published within two decades. This is the largest and the most comprehensive exploration of the computational data harmonisation strategies to the best of our knowledge. To provide a better scientific practice for the community working on data harmonisation, a comprehensive checklist with all the steps is proposed to guide the researchers on reporting their studies more effectively. With this checklist, explorations (what the strategy is) and advances (how well the model performs) of the study can be clearly illustrated by reporting the items in model and evaluation sections. Overall, the main contributions of this survey can be summarised as:

- A three-fold taxonomy that describes the methodology, evaluation and applications of computational data harmonisation strategies.
- A checklist with all the steps that can be followed in future data harmonisation studies.
- The critique and limitations of the existing data harmonisation strategies and potential studies.

The rest of the manuscript is organised as follows: (1) Section 2 describes the definition, motivation, utilisation and solution of computational data harmonisation issues; (2) Section 3 illustrates how this survey is conducted; (3) Sections 4,5, and 6 demonstrate the three-fold taxonomy of harmonisation strategies; (4) Section 7 describes the results of the meta-analysis and presents a checklist for data harmonisation studies; (5) Section 8 presents the checklist for harmonisation studies and summarises the critiques and limitations of current strategies; and (6) Section 9 concludes this survey.

## 2. Computational data harmonisation: definition, origin, what for and how?

This section illustrates the details of data harmonisation, including the definition, origin, purpose and solutions of computational data harmonisation tasks. To better describe these characteristics, the terminology of computational data harmonisation is illustrated in Table 2.

### 2.1. What?

Data harmonisation refers to combining the data from different sources to one cohesive data set by adjusting data formats, terminologies and measuring units [12]. It is mainly performed to address issues caused by nonidentical annotations or records of different operators or systems, which requires a standard protocol for manual adjustment. The conventional approach for data harmonisation is performed by manually setting rules or terms to integrate multicentre datasets from health

**Table 2**
Terminology of computational data harmonisation.

| Terminology | Definitions |
|---|---|
| Cohort | A group of data acquired by the same acquisition protocol and devices |
| Subjects | Patients (objects) involved in the study |
| Category | The classes that were involved in the study, e.g., cancer vs. normal |
| Cases | Samples (a subject can produce multiple samples with different acquisition protocols) involved in the study |
| Cohort bias | The non-biological related variances caused by acquisition protocols (also named as "batch effect" in gene expression studies) |
| Source cohorts | The cohort that needs to be harmonised from |
| Reference cohort | The cohort that needs to be harmonised to |

**Table 1**
Comparison of existing data harmonisation review studies.

| Survey | [6] | [7] | [8] | [9] | Ours |
|---|---|---|---|---|---|
| Period | ~2020 | ~2019 | ~2020 | ~2021 | ~2021 |
| # of reviewed studies | N/A | 23 | 49 | 42 | 96 |
| Domain | Radiomics | Radiomics | Radiomics | Radiomics | Radiomics, Gene, Pathology |
| Metric | × | × | × | × | √ |
| Checklist | × | × | × | × | √ |
| Guidance | × | √ | × | × | √ |
| Meta-analysis | × | √ | × | × | √ |

"# of reviewed studies" indicates the number of included papers in the survey.
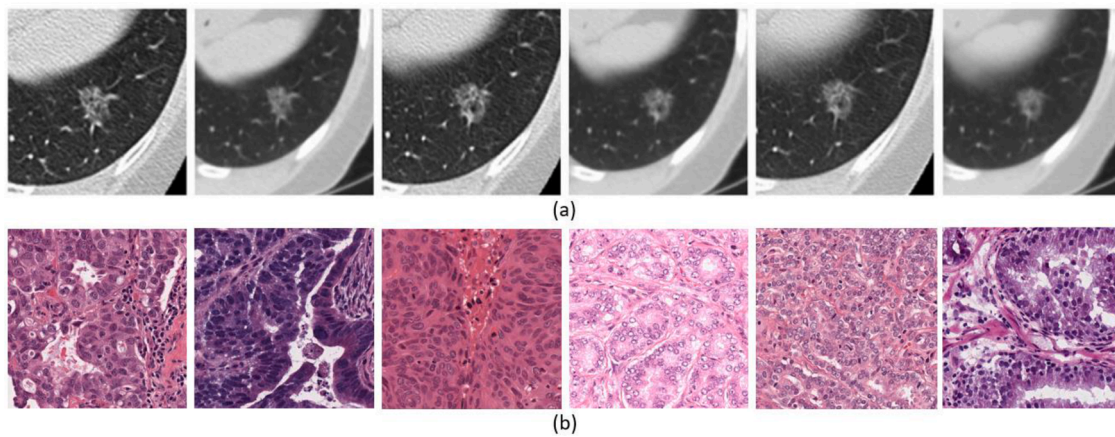
**Fig. 1.** Visualised differences in (a) radiomics and (b) pathology images. (a) a lung tumour captured on the same CT scanner with 6 different acquisition protocols (From [13]). (b) H&E stained tissue images from different sites [14].

information systems. It requires complex mapping of terminologies and manual harmonisations.

Different from manual harmonisation that relies on a standard protocol and manual adjustment, computational data harmonisation in digital healthcare aims to reduce the cohort bias (non-biological variances) given by different data acquisition schemes. It applies

**Table 3**

Summary of the reproducibility/repeatability studies.

| Reference | Intra-repro | Inter-repro | Repeatability | Condition | Variables | Object | Modality |
|---|---|---|---|---|---|---|---|
| Jha et al. [23], 2021 | 30.7% (332/1080) | 14.3% (154/1080) | 82.2% (888/1080) | ICC >0.90 | Slice Sickness | Phantoms | CT |
| Emaminejad et al. [24], 2021 | 8.0% (18/226) | / | / | CCC>0.90 | Reconstruction | Patients | CT |
| | 7.5% (17/226) | / | / | CCC>0.90 | Radiation Dose | Patients | CT |
| Kim et al. [25], 2021 | 11.0% (112/1020) | / | / | CCC>0.85 | Acceleration Factors | Patients | MRI |
| Ymashita et al. [26], 2020 | / | 5.6% (15/266) | / | CCC>0.90 | Different Scanners | Patients | CECT |
| Fiset et al. [27], 2019 | / | 22.6% (398/1761) | / | ICC >0.90 | Different Scanners | Patients | MRI |
| Saeedi et al. [28], 2019 | 20.5% (8/39) | / | / | CoV< 5% | Tube Voltage | Phantoms | CT |
| | 30% (13/39) | / | / | CoV< 5% | Tube Current | Phantoms | CT |
| Meyer et al. [29], 2019 | 20.8% (22/106) | / | / | $R^2 > 0.95$ | Radiation Dose | Patients | CT |
| | 52.8% (56/106) | / | / | $R^2 > 0.95$ | Reconstruction | Patients | CT |
| | 39.6% (42/106) | / | / | $R^2 > 0.95$ | Reconstruction | Patients | CT |
| | 12.3% (13/106) | / | / | $R^2 > 0.95$ | Slice Sickness | Patients | CT |
| Perrin et al. [30], 2018 | 24.8% (63/254) | / | / | CCC>0.90 | Injection Rates | Patients | CECT |
| | 13.4% (34/254) | / | / | CCC>0.90 | Resolution | Patients | CECT |
| Midya et al. [31], 2018 | 11.7% (29/248) | / | / | CCC>0.90 | Tube Current | Phantoms | CT |
| | 19.8% (49/248) | / | / | CCC>0.90 | Noise | Phantoms | CT |
| | 63.3% (157/248) | / | / | CCC>0.90 | Reconstruction | Patients | CT |
| Altazi et al. [32], 2017 | 21.5% (17/79) | / | / | Mean difference <25% | Reconstruction | Patients | PET |
| Zhao et al. [13], 2016 | 11.2% (10/89) | / | / | CCC>0.90 | Reconstruction | Patients | CT |
| | / | / | 69.7% (62/89) | CCC>0.90 | / | Patients | CT |
| Hu et al. [33], 2016 | / | / | 64.0% (496/775) | ICC>0.80 | / | Patients | CT |
| Choe et al. [34], 2019 | 15.2% (107/702) | / | / | CCC>0.85 | Reconstruction | Patients | CT |

CCC: concordance correlation coefficient; ICC: intraclass correlation coefficient; CoV: coefficient of variation; $R^2$: R-squared; CT: computed tomography; MRI: magnetic resonance imaging; CECT: consecutive contrast-enhanced computed tomography; PET: positron emission tomography.

computational strategies (such as machine learning, image/signal processing) to integrate multicentre datasets and reduce their non-biological heterogeneity. Compared with data cleansing, data normalisation, standardisation, etc., data harmonisation has a broader definition and is a term that represents the strategies of reducing cohort biases (caused by different acquisition protocols and devices). It can be conducted by removing outliers (data cleansing), aligning the location-and-scale parameters of cohorts (data normalisation), converting multiple datasets into a common data format (data standardisation/transformation, referring to manual harmonisation). It is of note that data harmonisation is not as same as style transformation (e.g., generating T1 images using T2 in MRI, or generating CT using X-Ray images), it only focuses on intra-modality datasets.

### 2.2. Why?

This section first illustrates the motivation of computational data harmonisation approaches, then describes the source of non-biological variances. Computational methods refer to the automatic analysis of digital healthcare data, using machine learning or mathematical modelling algorithms. It usually requires the extraction and fusion of data-derived features from the raw data. For instance, the grey level co-occurrence matrix (GLCM), which is one of the most commonly used textual features in radiomics, can be used as an independent prognostic factor (representing in F-FDG PET/CT images the metabolic intra-tumoral heterogeneity) in patients with surgically treated rectal cancer [15]. However, datasets acquired from different sites present significant variances (Fig. 1), which can hinder the effectiveness of extracted features and lead to unstable performance for both computational and manual diagnosis. In particular, Zhao et al. [13] found a considerable segmentation based inconsistency of lung tumours while conducting repeated manual labelling by three radiologists. This inconsistency could lead to a significant reduction (from 0.76 to 0.28) of concordance correlation coefficients for certain radiomics features. Therefore, computational data harmonisation is proposed to eliminate or reduce these non-biological variances in multicentre datasets for (1) enhancing the robustness and reproducibility of computational modules; (2) producing the fusion of knowledge captured beforehand with knowledge captured over a new task; (3) promoting the comprehensive performance of computational modules.

The non-biological data variances are mainly from hardware (e.g., scanners and platforms), acquisition protocols (e.g., signal/imaging acquisition parameters) and laboratory preparations (e.g., staining and slicing). These variances may lead to the weak reproducibility of quantitative biomarkers and limit the time-series studies based on multi-source datasets, indicating an urgent need for data harmonisation strategies to generate reproducible features [15,16].

#### 2.2.1. Heterogeneity of acquisition devices (inter-device variability)

Heterogeneity of acquisition devices leads to the variance of multicentre data, which is mainly discovered in signals, CT, MRI, and pathological images. This heterogeneity is mainly brought by different detector systems of vendors, the sensitivity of the coils, positional and physiologic variations during acquisition, and magnetic field variations in MRI, amongst others. [17–20]. Studies have shown that even using a fixed acquisition protocol for different brands of scanners, some radiomics features are still non-reproducible. For instance, Berenguer et al. [21] explored the reproducibility of radiomics features on five different scanners with the same acquisition protocol and witnessed large differences, ranging from 16% to 85% of the radiomics features were reproducible. Sunderland et al. [22] explored the large variance of standard uptake value (SUV) in different brands of scanners, witnessing a much higher maximum SUV of newer scanners compared with old ones.

#### 2.2.2. Heterogeneity of acquisition protocols (intra-device variability)

The different acquisition protocols are the main reasons for cross-cohort variability. They mainly include the scanning parameters (e.g., voltage, tube current, the field of view, slice thickness, microns per pixel, etc.) and reconstruction approaches (e.g., different reconstruction kernels) [35]. To investigate the intra/inter reproducibility of radiomics features, several studies have been conducted by test-reset experiments (Table 3). In Table 3, a good reproducibility/repeatability is defined as the high correlation coefficient (e.g., ICC, CCC, $R^2$) or low difference (e.g., mean difference, CoV) between two features. For instance, a certain radiomics feature is considered reproducible/repeatable when the CCC between features extracted from two repeated scans is larger than 0.90. As shown in Table 3, the scanning parameters notably affect the radiomics features, making the statistical analysis difficult. For instance, only 15.2% of radiomics features are reproducible when using soft and sharp kernels during the reconstruction [34]. This weak reproducibility greatly hinders the large-scale digital healthcare studies and applications of computational models. Although implementing strict standard protocol can reduce non-biomedical variances, the non-standard acquisition protocol is needed by physicians for personalised centre-based image quality considerations. For instance, the thickness and pixel size are regularly adjusted on a case-by-case principle to improve the data quality [36]. Therefore, the heterogeneity of acquisition protocol is unavoidable which requires a general solution.

#### 2.2.3. Heterogeneity of laboratory preparations (Preparation variability)

All the gene expression, radiomics, and pathological data heavily suffer from laboratory variances, including sample preparation, assay, slicing, and staining. For single-cell RNA sequencing (scRNA-seq) and microarray data, there are various analysis platforms with different biases, making it difficult to integrate and compare results from multi-centre/batch of data [37,38]. For radiomics data, variances such as injection rate and radiation dose may also affect the data quality. Considering the pathology data, variances are mainly from manual operations [39,40] (e.g., biopsy sectioning, sample fixation, dehydration and stain concentration), all these factors result in the variation of pixel values and stain consistencies.

### 2.3. What for?

**Large scale and longitudinal studies.** The challenges of integrating and utilising multicentre datasets make researchers realise the importance of data harmonisation when conducting large-scale studies [41]. On the one hand, the information fusion without harmonisation cannot achieve reproducible results in large scale and longitudinal studies [13, 31,42]. Some researchers have advised that the conclusions reached must be treated with caution since some features can vary greatly against minor non-biomedical changes [43]. Data harmonisation, on the other hand, is critical for patients who are monitored longitudinally and imaged on different scanners. For instance, the longitudinal PET cannot provide helpful information if they are gathered from multi-scanners, since the relationship between SUV and outcomes may get concealed [16].

**Transferability of computational models.** The unstable performance has been found when applying computational models to multi-centre datasets [44]. To address this issue, transfer learning was proposed to enhance the robustness of computational models by holding a priori knowledge on the way data can vary. It feeds the model with further data which reflects the variability that the model may encounter at inference time. However, transfer learning requires extra training samples to reduce the uncertainty with respect to the variability of data that models can cope with. This could be inapplicable for prospective studies in the digital healthcare field. Different from transfer learning, computational data harmonisation strategies can process the data without extra training or fine-tuning, which provide an applicable solution for multicentre studies. Meanwhile, there has been mounting
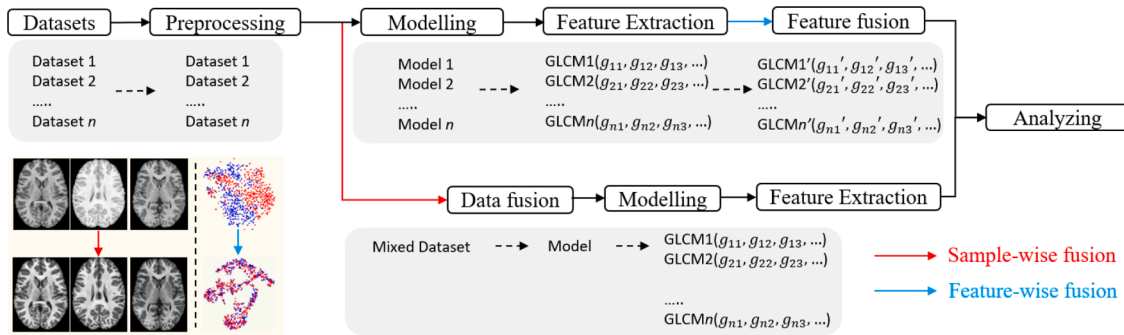
**Fig. 2.** Workflow of developing a computational data harmonisation method.

evidence that combining data harmonisation with machine learning algorithms enables robust and accurate predictions on multicentre datasets [45].

### 2.4. How?

The deployment of a computational method includes preparation (acquiring datasets such as staining, scanning), pre-processing, modelling and analysing, while the data harmonisation can be performed through the processing of images/signals/gene matrices (i.e., sample-wise) or alignment of data-derived features (i.e., feature-wise). The sample-wise harmonisation is usually conducted before modelling, aiming to reduce the cohort variance of all training samples and fuse multicentre samples as a single dataset. It involves image processing, synthesis and invariant feature learning approaches. After acquiring cohort-invariant data, a single model can be developed for clinical related tasks. The feature-wise harmonisation aims to reduce the bias of extracted features, such as the GLCM, convex hull area of the region of interest. It is usually performed on extracted feature matrixes, eliminating the cohort variances through fusing the extracted features (shown as the left bottom subfigure Fig. 2, the red and blue dots indicate samples from different cohorts). Both the sample-wise and feature-wise data harmonisation can effectively reduce the variances and improve the performance of the analysis. However, the feature-wise harmonisation requires several models to extract features of interest, leading to complex model development. Moreover, when the number of samples in each cohort is small, it is hard to develop the corresponding models.

### 3. Methods

#### 3.1. Literature search and review

The literature search, selection and recording were conducted independently by two researchers with experience in computer science and biomedicine. The agreement was then achieved by a third reviewer with the expertise of biomedical data analysis. All these searches were performed on Scopus Preview (Elsevier) database for publications up to July 10, 2021. To investigate the strategies of harmonisation for information fusion, we searched the literature using the keyword of 'batch effect removal', 'deep learning' and 'harmonisation', 'data
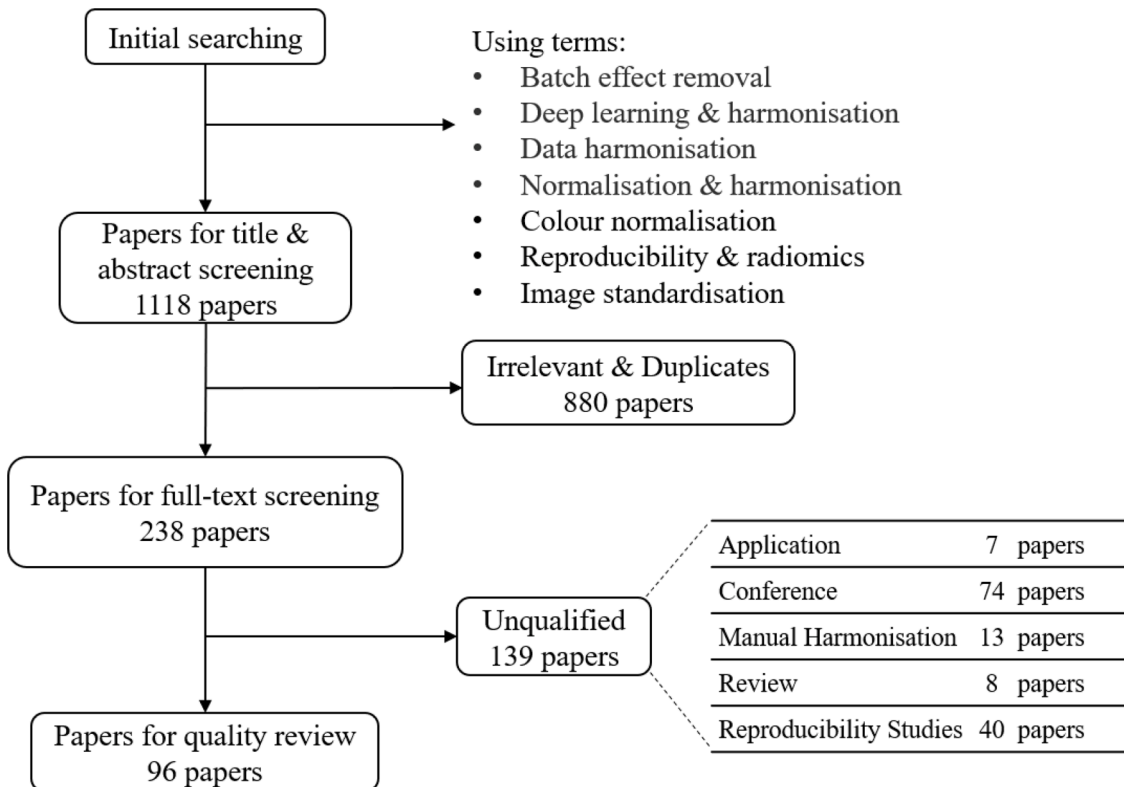


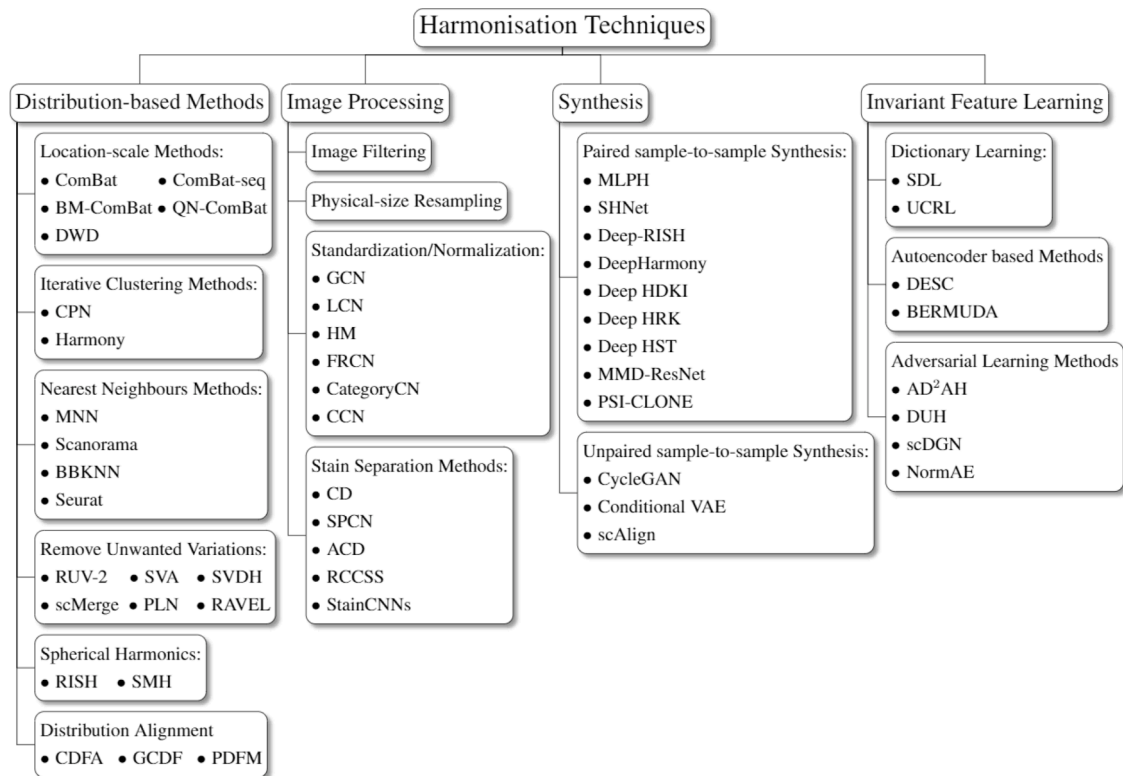**Fig. 3.** Literature selection procedure.

**Fig. 4.** Taxonomy of computational data harmonisation strategies.

harmonisation', 'normalisation' and 'harmonisation', 'colour normalisation', 'reproducibility' and 'radiomics', 'image standardisation'. These initial keywords were searched both independently and jointly to cover more literature. It is of note that both 'normalisation' and 'standardisation' are methods of harmonisation. The pre-screening was first conducted by viewing the abstract and title to filter those irrelevant articles. The eligibility was then checked through our criteria (given in Section 3.2) to remove the unqualified works for full-text review.

A flowchart demonstrating the literature selection procedure is presented in Fig. 3. After removing the irrelevant and duplicated articles by screening the titles and abstracts, 238 articles were selected for full-text screening. Based on eligibility criteria, 139 publications were considered unqualified, and 96 papers were included in this systematic review.

### 3.2. Inclusion and exclusion criteria

The entry criteria were: (1) original research publications in peer-reviewed journals or international conferences; (2) focus on the computational data harmonisation of digital healthcare data. The excluded criteria were: (1) studies that only applied existing harmonisation strategies without further development; (2) studies that focused on manual harmonisation such as regulations; (3) review and literature survey studies; (4) studies that only explore the reproducibility or stability without developing harmonisation approaches.

### 3.3. Data collection

Details of papers for quality review were manually summarised in a spreadsheet, including title, modality, methodology, metrics, data scale, year of publication, data property (e.g., private or public), applications, number of cohorts, and number of cases.

### 4. Data harmonisation strategies for information fusion

In this systematic review, data harmonisation approaches were divided into four groups, with the distribution based methods, image processing, synthesis, and invariant feature learning. To better illustrate the basic idea and relationship of computational approaches, a taxonomy is shown in Fig. 4, followed by a detailed description of harmonisation techniques.

### 4.1. Distribution based methods

The distribution based methods estimate/calculate the bias between cohorts from the latent space, then match/map the source data to the target ones through a bias correction vector or alignment functions.

#### 4.1.1. Location-scale methods (LS)
The location-scale methods estimate the location-scale parameters (mean and variance) of each cohort and align all data towards the same location-scale.

**ComBat**: ComBat [46] robustly estimated both the mean and the variance of each batch using empirical Bayes shrinkage, then harmonised the data according to these estimates. The data was first standardised to have similar overall mean and variance, followed by the empirical Bayes estimation via parametric empirical priors. With these adjusted bias estimators, the data could be harmonised by the location-scale model based functions [47 66]. For instance, Radua et al. applied ComBat to address the heterogeneity of cortical thickness, surface area and subcortical volumes caused by various scanners and sequences [53]. Whitney et al. implemented ComBat to harmonise the radiomic features extracted across multicentre DCE-MRI datasets [54].

**ComBat-seq**: Researchers have made more extensions based on the original ComBat harmonisation. Since the assumption of Gaussian distribution in the original ComBat made it sensitive to outliers, Zhang et al. proposed ComBat-seq [67] by assuming the Negative Binomial distribution, which could better address the outlier issues. The

ComBat-seq first built a negative binomial regression model and obtained the estimators of cohort bias, followed by the calculation of 'batch free' distributions for mapping original data.

**BM-ComBat:** Different from the original ComBat that shifted samples to the overall mean and variance, an M-ComBat [68] was proposed to provide a flexible solution, transferring the data to the location and scale of a pre-defined "reference". With these efforts, Da-ano et al. [69] proposed a BM-ComBat by introducing a parametric bootstrap in M-ComBat for robust estimation, aiming to provide a more flexible and robust harmonisation strategy.

**QN—ComBat:** Müller et al. [70] applied a quantile normalisation before ComBat correction in longitudinal gene expression data to achieve better performance.

**Distance-Weighted Discrimination (DWD):** DWD [71] searched the hyperplane where the samples could be well separated and projected the different batches on the DWD plane. The data was then harmonised by subtracting the DWD plane multiplied by the batch mean. It is of note that DWD repeated the translations of samples from different cohorts until their vectors were overlapped.

*4.1.2. Iterative clustering methods (IC)*

The iterative clustering methods harmonise the cohort bias by conducting multiple bias correction through repeated clusterings procedures. These methods usually (1) perform cluster to all samples from different cohorts, and (2) compute the correction vectors for harmonisation based on cluster centroids.

**Cross-platform normalisation (XPN):** XPN [72] took the combined standardised sample and median central gene as input to remove gross systematic differences, followed by the clusters, aiming to identify homogenous groups of genes and samples with similar expressions in combined data. The gene clusters were then acquired by assignment function, which was used to compute estimated model parameters via standard maximum likelihood.

**Harmony:** Harmony [73] first employed principal components analysis (PCA) to reduce the dimension of all samples, and classified them into several groups (one centroid per group) through k-means clustering. With these centroids, the correction factors for harmonisation were calculated. The above clustering and correction were repeated until the convergence.

*4.1.3. Nearest neighbours methods (NNM)*

NNM methods first found the mutual nearest pairs, then computed the bias correction vectors based on paired samples and subtracted these vectors from the source cohort. Differences in these methods mainly refer to the geometry space when locating the mutual nearest pairs.

**Mutual nearest neighbours (MNN):** MNN identified nearest neighbours between different cohorts and treated them as anchors to calculate the cohort bias [74]. It first pre-normalised the gene data with cosine normalisation, followed by the estimation of the bias correction vector by computing the Euclidean distances between paired samples. The bias correction vector was then applied to all samples instead of the participated pairs. It required that all participated batches must share at least one common type with another.

**Scanorama:** Similar to the MNN method, panorama stitching (Scanorama) [75] aims at estimating cohort bias from samples across batches. It first reduced dimensions of raw data (or source data) using singular value decomposition (SVD). Then an approximate nearest neighbour was adopted to find the mutually linked samples across cohorts. Different from MNN, Scanorama checked the priority of dataset merging within all batches and acquired the merged panorama based on the weighted average of batch correction vectors. At last, the harmonisation was performed with Scanpy [76] workflow.

**Batch balanced k-nearest neighbours (BBKNN):** Initially, BBKNN [77] found the nearest neighbours in a principal component space based on Euclidean distances. Then it built a graph that linked all the samples across cohorts based on the neighbour information. These neighbour sets were then harmonised by uniform manifold approximation (UMAP) [78] algorithms.

**Standard CCA and multi-CCA (Seurat):** Different from other NNM-methods, Seurat [79] performed canonical correlation analysis to acquire the canonical correlation vectors that could project multi-datasets into the most correlated subspace. In this subspace, the mutual nearest pairs were located to compute the bias correction vectors to guide the data integration. When processing multi-cohort datasets (number of cohorts larger than two), the first batch would be set as the reference batch for the correction of the second batch. Then the harmonised second batch would be appended to the reference batch. This repeated procedure stopped when all the batches are harmonised [38, 79].

*4.1.4. Remove unwanted variations (RUV)*

These methods assumed that the cohort bias was independent of those biases refer to biological variances, which could be estimated as "unwanted variations". For instance, the bias of negative control genes (prior known genes that would not be affected by biological changes of interest) could be regarded as cohort bias. Based on this assumption, the raw data could be harmonised by subtracting those "unwanted variations".

**Remove unwanted variations, 2-step (RUV-2):** Control variables were used by RUV-2 to discover the factors related to cohort bias [80]. The negative control (probes that should never be expressed in any sample) samples were subjected to component analysis, and the resulting factors were incorporated into a linear regression model. Variations in the expression levels of these genes thus were considered undesirable. To extract low-dimensional features, Risso et al. [81] presented an extension of the RUV-2 with a zero-inflated negative binomial model that accounted for dropouts, discretisation, and the count character of the data. The cohort bias was then subtracted from the raw data to generate a gene expression matrix that is harmonised.

**Singular value decomposition harmonic (SVDH):** By factorising the expression matrix of input data and reconstructing it while taking off the elements related to the cohort bias, singular value decomposition (SVD) could be used to reduce cohort bias. Alter et al. [82] suggested using SVD to harmonise the data by filtering away the eigenarrays that lead to noise or experimental artefacts.

**scMerge:** scMerge [83] first constructed a graph that connected clusterings between cohorts by searching for mutual nearest neighbours. The unwanted factors were then estimated using stably expressed genes as negative controls. At last, an RUV model was used to collect and remove unwanted differences between cohorts.

**Surrogate variable analysis (SVA):** SVA [84] aimed to recognise and estimate the unwanted variations of data from multiple cohorts. It could be performed without any cohort information. The mixed dataset was first divided into a collection of $n$ surrogate variables via SVD, followed by the clearance of data with large variances. SVA coefficients were then calculated for harmonisation by using a linear regression function with surrogate variables and raw diffusion intensities.

**Print-tip loess normalisation (PLN):** PLN [85] was initially proposed to deal with microarray data. To eliminate the cohort bias, PLN employed a blocking term to construct a linear model with the input data. The cohort bias was subtracted from the original data to produce the batch corrected expression matrix.

**Removal of artificial voxel effect by linear regression (RAVEL):** RAVEL [86] separated the voxel value into unwanted variation parts and biological parts. The unwanted variation factors were estimated from the region of interest by SVD, based on the prior knowledge of voxel values, which were not related to disease status.

*4.1.6. Spherical harmonics (SH)*

Spherical harmonics approaches were designed to harmonise MRI data, aiming to coordinate all data from different cohorts to the same spherical harmonic domain, by adjusting the spherical variables.

**Rotation invariant spherical harmonics (RISH):** RISH was based

on mapping diffusion-weighted imaging data from source cohorts to target cohorts [17,66,87,88]. It started with calculating the rotation-invariant features from the estimated spherical harmonics coefficients (of target and source samples, respectively). These rotation invariant features were then mapped from the source cohorts to target cohorts through region-specific linear mapping, followed by the updating of spherical harmonics coefficients. The harmonised diffusion signal was calculated for each subject in source cohorts using the latest spherical harmonics coefficients in target cohorts of gradient directions.

**Spherical moment harmonics.** Due to the insufficient adjustment by location-scale parameters in some cases, researchers proposed the spherical moment method (SMM), which utilised the spherical moments to map the diffusion-weighted images from source cohorts to reference cohorts [89,90]. SMM matches the spherical mean ($M_1$) and spherical variance ($C_2$) per b-value (the diffusion weighting) by $M_1[T_b] = M_1[f(S_b)]$ and $C_2[T_b] = C_2[f(S_b)]$, where $T_b, S_b$ are data from the target and source cohorts under $b$ shell, respectively. The mapping parameters for harmonising data from different cohorts were acquired by the linear transform $f$.

### 4.1.7. Distribution alignment (DA)

Distribution alignment methods aim to transform the distribution of the source cohort to that of the reference cohort, using cumulative distribution functions or probability density functions.

**Cumulative distribution functions alignment (CDFA):** CDFA [91] was first proposed for multisite MRI data harmonisation, which aligned the source voxel intensities through an estimated non-linear intensity transformation to match the target cumulative distribution functions. The estimated intensity transformation defined a one-to-one mapping between the voxels in source and target cohorts.

**Gamma cumulative distribution functions alignment (GCDF):** The voxel intensities were re-parameterised using a mixture model of two Gamma distributions that fitted a reference histogram [92]. This reparameterisation was based on the CDF of the Gamma component, which modelled the particular uptake, and constrained the new feature space to [0, 1].

**Probability density function matching:** GENESHIFT [93] estimated the empirical density and measured the distance between probability density functions. GENESHIFT first picked the common genes from different cohorts, then estimated their probability density functions to find the best matching offsets. The harmonised data would be acquired by subtracting the estimated offsets from the source cohorts.

### 4.2. Image processing

Image Processing employs digital image processing algorithms to harmonise multi-cohort data, including image filtering (also called image convolution), registration, resampling and normalisation.

### 4.2.1. Image filtering (IF)

Image filtering (also called convolution) is the process that multiplies two arrays to produce a new array of the same dimension. The 2D second-order Butterworth low-pass filter was found to be able to eliminate cohort bias between CT images with different voxel sizes [94], while the local binary pattern filtering could produce stable and reproducible radiomic features [95].

### 4.2.2. Physical-size resampling (Resample)

Studies have shown that physical size such as pixel/voxel size, mpp (microns per pixel of level 0 in digital pathology) can greatly affect the radiomic/pathological features. This bias can be reduced using bilinear resampling to equalise all the physical sizes [94].

### 4.2.3. Standardisation/normalisation (SN)

Standardisation/normalisation models were designed to reduce the variation and inter-variability in different cohorts by linear transform.

These methods usually performed location-scale shifts in image spaces (e.g., HSV, RGB, $\alpha\beta$, illumination spaces, etc.) or image histograms.

**Global colour normalisation (GCN)** transfers the colour statistics from the source to the target images by globally altering the image histogram [96,97]. A typical representative of GCN is Z-score normalisation, assumed the variable from cohort $i$, subject $j$ as $X_{ij}$, z-score normalisation is conducted through

$$X_{ij} = \frac{X_{ij} - \mu_i}{\sigma_i} \tag{1}$$

where $\mu_i$ and $\sigma_i$ are the mean and standard deviation of each cohort. However, this global alignment may lose some information.

**Local colour normalisation (LCN)** transfers the colour statistics of the specific regions, e.g., ignoring the background regions, from source to target images. In [98], the authors first converted the source and target images from the RGB into the $l\alpha\beta$ space, and then conducted a transformation to harmonise the source image and re-converted it into the RGB space. It is of note that the luminance of background regions is not involved during the processing. This helped the transformation to preserve intensity information within the region of interest while requiring the pre-definition of certain regions.

**Histogram matching (HM):** HM is a method of contrast adjustment using the histogram of images [99]. It adjusts the distribution of images by scaling the pixel values to fit the range of specified histogram (i.e., the target one):

$$f(x, y) = \frac{I_{Tmax} - I_{Tmin}}{I_{Smax} - I_{Smin}}(I_S - I_{Smin}) + I_{Tmin} \tag{2}$$

where $I_T$ indicates the target image and $I_S$ is the source image. Generally, $I_{Tmax}$ and $I_{Tmin}$ are 0 and 255, respectively. For instance, Shah et al. [100] investigated the histogram normalisation on MRI images to harmonise cross-cohort data for multiple sclerosis lesion identification.

**Fuzzy based Reinhard colour normalisation (FRCN):** To decrease the colour variation, Roy et al. [101] applied fuzzy logic to regulate the contrast enhancement in $l$ space to adjust the colour coefficients within the $\alpha\beta$ space.

**Category based colour normalisation (CategoryCN):** To reduce the variance of global colour normalisation, researchers proposed a category based approach for accurate colour normalisation [102]. CategoryCN first classified each pixel by unsupervised approaches from the source and target images, then conducted colour normalisation based on the different classes.

**Complete colour normalisation (CCN):** The complete colour normalisation included the normalisation of illumination and spectrum, one to harmonise the illuminant during imaging and another to reduce spectral variation [39,103]. CCN estimated the illuminant and spectral matrices from the target cohort, then matched the source illuminant and spectral estimations to the target ones.

### 4.2.4. Stain separation methods (SS)

Stain separation approaches separated the input images into distinct channels (e.g., the haematoxylin channel, eosin channel, and the background channel for H&E-stained images) to evaluate the stain feature matrix and match these features through certain operations from source to target cohort data. The core concept of stain separation was based on Lambert Beer's law [104] (in the RGB space, stain concentrations are nonlinearly dependant), shown as

$$I_C = I_0 e^{-OD_c} \tag{3}$$

where $I_0$ was the value of incident light, and $OD_c$ was the value of images in optical density (OD) space. Most stain separation methods aimed to factorise the OD values into two matrices as

$$OD_c = \log\left(\frac{I_0}{I_C}\right) = S * D \tag{4}$$

where *S* was the stain depth matrix and *D* was the stain colour appearance (SCA) matrix.

**Colour deconvolution (CD):** These approaches estimated the concentration of stains in pixel values and normalised the spectral variation in separated stains [105 108]. For example, estimation of the stain matrix was first given by evaluating the proportion of RGB channels within different cohorts, followed by colour deconvolution [106,107]. The inverse of the staining appearance matrix was multiplied with the optical density space intensity value to get normalised stain channels using non-linear spline mapping.

**Structured-preserving colour normalisation (SPCN):** SPCN assumed that most tissue regions were characterised by the most effective stain amongst the used stains [109]. It first converted a given RGB image to optical density using the Beer-Lambert Law. After that, SPCN decomposed images into several stain density maps using sparse and non-negative matrix factorization (SNMF), followed by the combination of the stain density map and colour normalisation.

**StainCNNs:** Inspired by SPCN, Lei et al. proposed a deep neural network for stain separation to reduce the computational consumption of SNMF [110]. The proposed stainCNNs approach took the source images as input and learned to generate the stain colour appearance matrix. It significantly reduced the processing time while retaining the high quality of the harmonised images.

**Adaptive colour deconvolution (ACD):** ACD first transferred the input RGB images to optical density space, then performed stain separation with adaptive colour deconvolution matrix to obtain the haematoxylin (H) channel, eosin (E) channel and residual channel [111]. At last, the harmonised images were obtained through recombining the H and E components with a stain colour appearance matrix of target cohorts.

**Rough-fuzzy circular clustering based stain separation (RCCSS):** In RCCSS, stain separation was carried out using an image model based on transmission light microscopy [112]. Initially, each image was transferred to OD space and then decomposed to obtain the SCA matrix and associated stain depth matrix. Maji et al. [113] presented a circular clustering algorithm to find the 'centroid', 'a crisp lower approximation', and the 'fuzzy boundary', which could be integrated by saturation-weighted hue histogram in the HIS colour space.

### 4.3. Synthesis

The objective of synthesis is to precisely reproduce a sample that belongs to a missing modality or domain, which harmonises the multi-cohort datasets. It relaxes harmonisation tasks as style transfer and considers each cohort as a 'style' and transfers all samples to the same 'style'. Based on the characteristics of the training sample, synthesis methods are divided into paired synthesis and unpaired synthesis.

#### 4.3.1. Paired sample-to-sample synthesis (P-s2s)

P-s2s methods are trained using paired samples generated from the same object acquired using different protocols. These methods aim to learn the data transfer between source and reference cohorts, which require the repeated acquisition of the same subject under different protocols. Therefore, they can only be applied to radiomic data since the repeated acquisition for the same subject is impossible for gene expression and pathology.

**Multi-layer perceptron harmonic (MLPH):** In 2009, a pilot architecture of the autoencoder-related method was proposed by Cheng et al. [114] to generate the harmonised data by learning the nonlinear transform function.

**Spherical harmonic network (SHNet):** Golkov et al. [115] presented a cascaded fully connected network that employs ReLU and Batch normalisation to harmonise the diffusion MRI scans. Inspired by SHNet, Koppers et al. [116] applied the residual structure to improve the robustness while avoiding overfitting.

**Deep rotation invariant spherical harmonics (Deep-RISH):** Karayumak et al. [117] proposed a deep learning based non-linear mapping approach that utilises RISH features to map the raw signal (dMRI data) between scanners with the same fibre orientations. Deep-RISH was composed of five convolution layers, which took the $9 \times 9$ RISH feature patches as the input.

**DeepHarmony:** DeepHarmony was proposed to produce data with consistent contrast within different cohorts [118]. It employed a U-Net based architecture, taking data from the source cohort and producing harmonised data of the target cohort.

**Deep harmonics for diffusion kurtosis imaging (Deep HDKI):** Tong et al. [119] carried out a concise architecture with three 3D-convolution layers for diffusion kurtosis images (DKI). The paired data was generated using an iterative technique called linear least square and were non-linearly registered to diffusion-weighted images acquired on the target scanner using the computational tools. Then the neural network was trained on the paired samples for harmonisation.

**Deep harmonics for slice thickness (Deep HST):** Park et al. [120] studied the reproducibility of radiomic features in lung cancer under different slice thicknesses and proposed an end-to-end deep neural network to generate harmonised CT data between 1-, 3-, and 5-mm slice thickness.

**Deep harmonics for reconstruction kernel (Deep HRK):** Choe et al. [34] explored the influence of different reconstruction kernels on radiomic features and presented a CNN with residual learning to transfer the data from the soft kernel (B30f) to the sharp kernel (B50f).

**Distribution-matching residual network (MMD-ResNet):** Shaham et al. [121] presented a comprehensive multi-layer perceptron for harmonisation with residual connection [122] and batch normalisation [123] techniques. Given two cohorts of data $X [x_1, x_2, ..., x_m] \in D_1$ and $Y [y_1, y_2, ..., y_n] \in D_2$. The MMD-ResNet aimed to learn a map $\widehat{\varphi} : R_d \rightarrow R_d$ by minimising the maximum mean discrepancy [124] between $\widehat{\varphi}(X)$ and $Y$. It is of note that this was a 'one-way street' distribution matching for harmonisation and required re-training for inverse transformation.

**Pulse sequence information based contrast learning on neighbourhood ensembles (PSI-CLONE):** PSI-CLONE [125] first calculated sequence parameters $\varnothing_s$ from source cohorts, then applied $\varnothing_s$ to the reference cohorts to produce the source-style data. By training a regression model to learn the nonlinear mapping between synthesised source-style data and reference data, the source cohorts could be harmonised effectively. Based on PSI-CLONE, Jog et al. [126] applied the multi-scale feature extraction to improve the performance.

#### 4.3.2. Unpaired sample-to-sample synthesis (Up-s2s)

Up-s2s approaches generate the harmonised data by cycle-consistent generative adversarial networks or conditional variational autoencoder-decoder, which require sufficient samples and cohort labels from different cohorts for network training.

**Cycle-consistent generative adversarial networks (CycleGAN):** Most synthesis methods of unpaired sample-to-sample translation were based on CycleGAN [127,128] and its derivatives [62,129,130]. In [130], a CycleGAN with Markovian discriminator was applied to harmonise the diffusion tensor data, which was designed to further improve the ability to capture local information.

**Conditional variational autoencoder-decoder (Conditional VAE):** Variational Autoencoder (VAE) is commonly used in data synthesis, dimensional reduction, and feature refinement tasks. It employs an encoding network $E_\theta(z|x)$ to decompose the input high dimensional data $x$ into hidden representation $z$, and a decoding network $D_\delta(x|z)$ to reconstruct the raw data $x$, where $\theta$ and $\delta$ are parameters of $E$ and $D$. The conditional VAE modifies the decoder to a conditional decoder $D_\delta(x|z, c)$ that takes the latent variable $z$ and specified cohort $c$ back to a harmonised data $\widehat{x}$. By integrating Conditional VAE with the adversarial module, cohort transfer can be performed without paired training samples. Several studies have been proposed using Conditional VAE for data harmonisation, including:
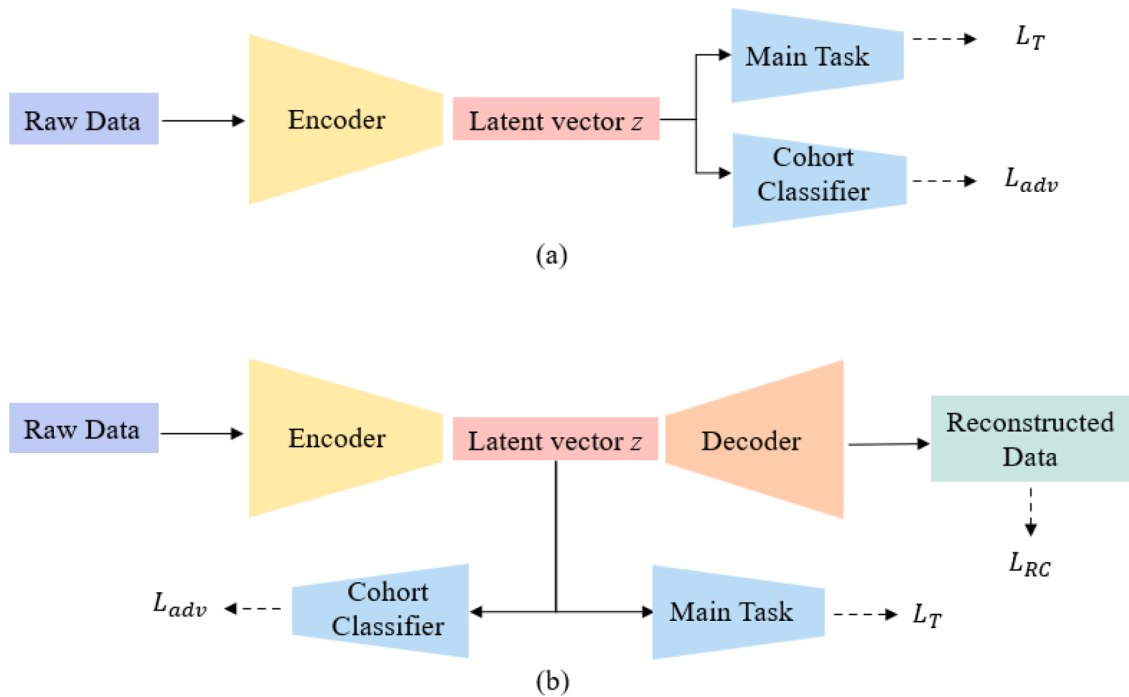
**Fig. 5.** Illustration of adversarial learning methods.

(1) **SH-VAE** [131] performed cohort bias correction of diffusion-weighted MRI by conditional VAE to produce cohort-invariant encodings. Different from other conditional VAE based methods, SH-VAE took spherical harmonics coefficients as input and output.

(2) **stVAE** [132] applied Conditional VAE with Y-Autoencoders (additional classification head in the encoder) and adversarial feature decomposition for single-cell RNA sequencing.

(3) **scAlign** [133] performed harmonisation by learning a duplex mapping of cell sequences between different cohorts in a low dimensional latent space. This mapping enabled the model to estimate a representation of certain samples under data from different cohorts. Besides, it employed the "association learning" method [134] to walk through the embeddings generated by a neural network with data from different cohorts. The association learning enabled the network to extract the embeddings that can capture the essence of the input data, leveraging the lack of annotations in paired synthesis. With these essence embeddings, scAlign applied a decoder to synthesise the harmonised data.

(4) **iMAP** [135] first presented an autoencoder architecture to learn the cohort-invariant features, then used these features to set MNN pairs by the random walk strategy. This autoencoder included one encoder $E$ and two generators $G_1$ and $G_2$, with two inputs (gene expression vectors and cohort labels) and outputs ($G_1$ for generating the cohort variations and $G_2$ for reconstructing the original input), respectively. With the defined MNN pairs, a GAN model was used to produce the cohort-invariant samples.

### 4.4. Invariant feature learning

The invariant feature learning techniques are meant to learn the cohort-invariant features from different cohorts of data, then apply these features for the main task (e.g., segmentation, classification, regression). The concept behind representation learning approaches for harmonisation is that if a sparse dictionary/mapping can be built from data of different cohorts, these learnt representations will not include inter/intra cohort variability.

#### 4.4.1. Dictionary learning (DictL)

**Sparse dictionary learning (SDL):** SDL [136][137] was a representation learning approach that aimed to reduce the complexity of the harmonisation task by decomposing the input data as a linear combination of components. SDL could be applied to identify the cohort-invariant features to reconstruct the raw data from a huge number of random features [170].

**Unsupervised colour representation learning (UCRL):** UCRL [138] first estimated the sparsity parameter based on SPCN, then employed a robust dictionary learning method [139] to acquire the stain colour appearance matrix. By taking the stain centroid estimation as an $L_1$-regularised linear least-squares task, the stain mixing coefficients map of the source data was combined with the colour appearance matrix of the reference data.

#### 4.4.2. Autoencoder based methods (AE)

**DESC:** DESC [140] trained a VAE to obtain the cohort-invariant feature embeddings, then iteratively optimised a clustering loss function to group the cohort data. The Louvain clustering [141], which aimed to improve modularity for community detection, was used to initialise the cluster centres.

**BERMUDA:** BERMUDA [142] first applied a graph based clustering to data from different cohorts individually, followed by a method (named MetaNeighbor) to identify similar clusters between cohorts to get the initial unaligned comprehensive dataset. An autoencoder was then built to reconstruct the input data while producing invariant feature embeddings in the low dimensional latent space. These feature embeddings were cohort-invariant and can be used for further analysis.

#### 4.4.3. Adversarial learning methods (AdvL)

The adversarial learning methods indicate developing a learning system that focuses on the scanner/protocol invariant features while simultaneously maintaining performance on the main task of interest, thus reducing the cohort bias on predictions. These methods [143 146] were usually composed of an adversarial module for cohort identification, a backbone for feature extraction, and the main task for classification, regression, and/or segmentation.

The adversarial learning methods used for harmonisation mainly had
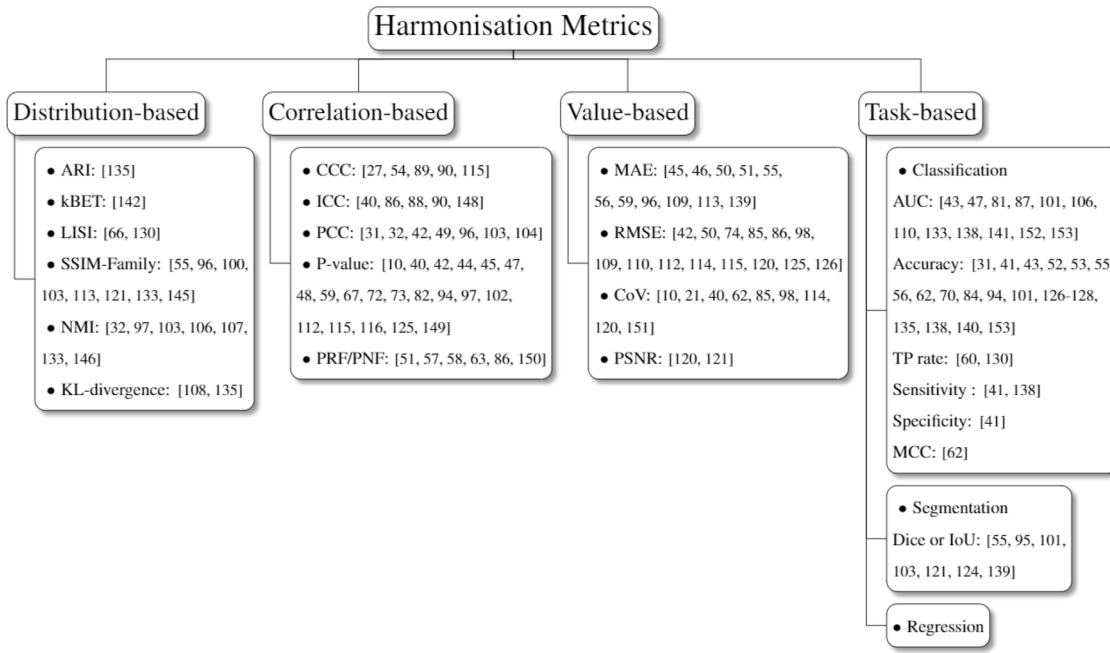
**Fig. 6.** Taxonomy of harmonisation metrics. The visualization and cohort classification assessment are not presented due to their limited subcategory.

two structures, as shown in Fig. 5(a) and (b). For methods such as AD$^2$AH (Attention-guided deep domain adaptation harmonics) [143], DUH (Deep unlearning harmonics) [144], and scDGN (single-cell domain generalisation network) [145], the adversarial module (domain classifier) was designed to assist the encoder to learn the cohort invariant features by maximising the adversarial loss $L_{adv}$ while minimising the loss of the main task $L_T$ (Fig. 5(a)). To acquire the precise feature representation $z$, methods such as NormAE [146] added a decoder to reconstruct the input raw data through minimising the reconstruction loss $L_{RC}$, shown in Fig. 5(b). By incorporating these optimisation functions, the main task could achieve stable performance when dealing with multi-cohort data.

## 5. Evaluation approaches of the data harmonisation strategies

This section explores evaluation approaches for harmonisation performance and divides them into distribution based, correlation based, value based, and task based metrics (Fig. 6). Distribution based metrics evaluate the harmonisation performance through assessing the clusters or location-scale parameters amongst different cohorts. The correlation based and value based metrics assess the variability of data-derived features from different cohorts to test the reproducibility. Besides, cohort classification is also considered as an evaluation method, aiming to demonstrate the harmonisation effect by cohort classification results. Visualisation is the commonly used evaluation approach that can straight visualise datasets before and after harmonisation.

### 5.1. Distribution based evaluation

Distribution based metrics assess the harmonisation performance via calculating the clustering or local-scale parameters. The clustering related metrics include adjusted rand index (ARI), k-Nearest neighbour batch-effect test (kBET), local inverse Simpson's index (LISI). The location-scale related metrics contain structural similarity families, normalised median intensity and KL divergence.

#### 5.1.1. Adjusted rand index (ARI)

The adjusted Rand index is the corrected-for-chance version of the Rand index (RI) and can be used for harmonisation evaluations [140].

Given a set of $n$ elements and their predictions $X = \{X_1, X_2, \ldots, X_i\}$ and $Y = \{Y_1, Y_2, \ldots, Y_j\}$, the RI can be calculated through

$$RI = \frac{TP + TN}{TP + FP + FN + TN},\tag{5}$$

where TP is the number of true positives, TN indicates the number of true negatives, FP is the number of false positives and FN is the number of false negatives. The ARI is illustrated as

$$ARI = \frac{RI - E[RI]}{\max[RI] - E[RI]},\tag{6}$$

where E(RI) is the expectation of the RI. It ranges from 0 to 1, and a large ARI indicates the cluster results are similar to the real labels.

#### 5.1.2. k-Nearest neighbour batch-effect test (kBET)

The k-Nearest neighbour batch-effect test was proposed to assess whether the distribution based harmonisation method can remove cohort bias while preserving biological variability [147]. kBET formulates a null hypothesis that the data is 'well mixed'. It employs a $\chi^2$ based test for random fixed-size neighbourhoods to evaluate whether the data is well mixed. The low average rejection rate indicates good harmonisation performance and vice versa. As a result, determining whether the mean rejection rate surpasses a significance level allows the null hypothesis to be rejected for the whole dataset.

#### 5.1.3. Local inverse Simpson's index (LISI)

LISI combines perplexity based neighbourhood construction and the Inverse Simpson's Index (ISI), which is sensitive to local diversity and can be well interpreted [73,135]. LISI applies the Gaussian kernel based distributions of neighbourhoods via distance based weights and computes the local distributions by fixing perplexity. Meanwhile, it uses the ISI to enhance the interpretation, that is

$$ISI = \frac{1}{\sum_{n=1}^{N} p(x)},\tag{7}$$

where $p(x)$ is the batch probabilities in local distributions.

### 5.1.4. Structural similarity families

Structural similarity index measure (SSIM) was designed to evaluate the image quality degradation during data transmission by measuring the similarity between two samples [148]. It was initially proposed for grey level images and has been widely applied to evaluate the harmonisation performance in digital pathology [62,101,108,118,126]. Assume $\mu_i$, $\sigma_i$ as the average and variance of sample $i$, $\sigma_{xy}$ as the covariance between sample $x$ and $y$, and the smooth parameters $c_1, c_2$. The SSIM can be described as

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{8}$$

To better assess the similarity for colour images, Kolaman et al. [149] proposed quaternion structural similarity (QSSIM) to measure the size and direction of chrominance, luminance and degradation [109]. Feature similarity index (FSIM) utilises phase congruency and gradient magnitude features to evaluate the low-level features of image visual quality [150]. The QSSIM and FSIM were employed in [110,138] and [105,110], respectively, to assess the structural preserving conditions after the harmonisation process. Though most methods applied structural similarity related metrics for evaluation, studies have shown their limitations and weaknesses [151]. For instance, it has been reported that these metrics suffer from uniform pooling, distortion and instability, especially when measuring samples with hard edges or low variance regions.

### 5.1.5. Normalised median intensity (NMI)

Assume the mean of R, G, B values for the i th pixel within the image $I$ as $A_i$, the NMI for assessing the colour consistency is calculated as

$$\text{NMI}(I) = \frac{\text{Median}(A_i)}{P_{95}(A_i)}, \tag{9}$$

where the $P_{95}$ denotes the 95th percentile [39,102,108,111,112,138,152]. The harmonisation strategy is effective when the median and maximum intensity values are close enough. Since the NMI does not consider the consistency of the ROI within the same biopsy set of $S$ images, Maji et al. [113] presented an extension of NMI, named Between-Image colour constancy (BiCC) index, which can be represented by

$$\text{BiCC}(I) = \frac{1}{2(|S|-1)} \times \sum_{J \neq I} \frac{\text{Median}(A_i) + \text{Median}(A_j)}{\max[\max(A_i), \max(A_j)]}, \tag{10}$$

where $i \in \text{ROI}(I)$ and $j \in \text{ROI}(J)$. The value of BiCC ranges from 0 to 1, an efficient harmonisation algorithm for image modality should make the value as high as possible.

### 5.1.6. Coefficient of variation (COV)

Given the mean $\mu$ and standard deviation $\sigma$, the coefficient of variation (COV) is defined as $\sigma/\mu$, which depicts the degree of variation in respect to the population mean [17,28,47,69,90,103,119,125,153]. The Multivariate COV (MCOV) is used to quantify the variability of features between different cohorts, with a lower value indicating better reproducibility [154]. Assume $\mu_x$ and $\mu_y$ as the mean of features extracted from two different cohorts $x$ and $y$, $\sum_{x,y}$ as the covariance matrix, the MCOV is computed via

$$\text{MCOV}(x,y) = \sqrt{\frac{\mu_{x,y}^T * \sum_{x,y} * \mu_{x,y}}{\left(\mu_{x,y}^T * \mu_{x,y}\right)^2}}, \tag{11}$$

### 5.1.7. Kullback-Leibler divergence

Kullback-Leibler (KL) divergence was proposed to measure how a probability distribution is different from another one. Assume two discrete probability distributions $p$ and $q$ (each with $k$ samples), the KL divergence is given by

$$\text{KL} = \sum p_k log \frac{p_k}{q_k} \tag{12}$$

It was applied as a metric for harmonisation strategies, with 0 indicating identical quantities of information between two distributions [113,140]. For instance, Li et al. [140] applied KL divergence to evaluate how randomly are samples from different cohorts mixed together within each cluster.

## 5.2. Correlation based evaluation

The measurement of reproducible/nonreproducible is usually given by statistical analysis via calculating the correlation between data-derived features before and after harmonisation, including concordance correlation coefficient, intra-class correlation coefficient etc.

### 5.2.1. Pearson correlation coefficient (PCC)

Pearson correlation coefficient measures the linear correlation between two groups of variables $X$ and $Y$, which is presented as

$$p_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{13}$$

where cov is the covariance and $\sigma$ indicates the standard deviation. PCC ranges from 0 to 1 where 0 denotes there is no correlation between $X$ and $Y$, and 1 represents a perfect correlation. The PCC was used as an indicator to assess the similarity between the source and harmonised data [38,39,49,56,101,108,109].

### 5.2.2. Concordance correlation coefficient (CCC)

The CCC was proposed by Lin et al. [155] that measured the agreement between two variables and was used to assess the reproducibility [34,61,94,95,120]. Different from PCC that can only assess the correlation between two groups of data, CCC measures how large the gap between two groups of data is. Assume the two variables $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_n\}$ and their mean $\overline{x}$, $\overline{y}$, and variance $s_x^2$, $s_y^2$, the CCC is given by

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\overline{x} - \overline{y})^2} \tag{14}$$

where $s_{xy} = \frac{1}{N} \sum_{n=1}^N (x_n - \overline{x})(y_n - \overline{y})$.

### 5.2.3. Intra-class correlation coefficient (ICC)

Both PCC and CCC can only assess the correlation between two groups of data, which cannot be implemented on multi-cohorts. The ICC is utilised for data structured as groups instead of those as paired observations, it is usually used to assess the variability within different protocols, different imaging devices, or different sites [47,91,93,95,156]. It interprets on a scale of [0, 1], with 1 illustrating the perfect agreement and 0 indicating complete randomness. Essentially, the ICC employed for data harmonisation describes the confidence of how similar the variables are in different cohorts, which is the one-way random model that assumes there is no systematic bias [47]. Data from various cohorts are pooled and assessed within or cross operators based on the analysis of variance (ANOVA). The one-way random model can be given from:

$$\text{ICC} = \frac{MS_B - MS_W}{MS_B + (n_G - 1)MS_W} \tag{15}$$

where $MS_B$ is the mean square between groups, $MS_W$ is the mean square within groups and $n_G$ indicates the average group size.

### 5.2.4. P-value

Some studies evaluate the harmonisation effectiveness through

computing the P-value given by paired hypothesis tests [17,47,49,51, 52, 54, 55,66,74,79,80,87,99,102,107,117,120,121,130,157]. This statistical analysis is often conducted based on the paired region of interests before-and-after harmonisation. In particular, there are significant differences (corresponding to p-value < 0.05) between the data of different cohorts before and after harmonisation, and vice versa. For instance, Fortin et al. [58] analysed the number of voxels that are significantly related to cohorts, e.g., a voxel is counted when the p-value calculated is less than 0.05.

### 5.2.5. Percentage of reproducible/nonreproducible features (PRF/PNF)

The percentage of nonreproducible features was treated as an evaluation metric [58,64,65,70,91,158], e.g., Mahon et al. [158] compared the percentage of significantly different features before and after conducting ComBat harmonisation. On the contrary, the percentage of reproducible features was also considered as an evaluation metric of harmonisation performance [61].

### 5.3. Value based evaluation

The value based evaluation mainly assesses the intensity differences between the data or data-derived features before and after harmonisation. This usually requires a "ground truth" that can ideally reflect harmonisation results, a low value of intensity differences indicates good harmonisation performance.

### 5.3.1. Mean absolute error (MAE)

The average absolute error of features (textual and clinical features) can be used to reflect harmonisation effects [52,53,57,58, 62, 63, 66, 101,114,118,144], this usually requires the extraction of certain ROIs from the data before and after harmonisation. For instance, Wachinger et al. [63] evaluated the MAE in age prediction on the raw dataset and ComBat-harmonised dataset to illustrate the effectiveness of ComBat. Dewey et al. compared the MAE between the synthesised and raw images to demonstrate the harmonisation performance.

### 5.3.2. Root-mean square error (RMSE)

Many researchers measured the RMSE between the harmonised samples and the ground truth targets to assess the replicability [49,57, 81,90, 91, 103, 114,115,117,119,120,125,130,131]. For instance, Moyer et al. employed RMSE and mean absolute error to assess the harmonisation performance between synthesised diffusion MRI and the ground truth. However, this metric requires paired datasets which is a heavy burden for digital healthcare research.

### 5.3.3. Peak signal to noise ratio (PSNR)

PSNR illustrates the ratio between the maximum power of a signal and the power of noise that influences the integrity of its representation. Consider two groups of variables $X$ and $Y$, the PSNR between $X$ and $Y$ can be given as

$$PSNR = 20 \times log_{10} \left( \frac{MAX_i}{\sqrt{MSE(X, Y)}} \right), \tag{16}$$

where $MAX_i$ denotes the maximum possible value of the image (255 for images) and MSE is the mean squared error. It is commonly used in image denoising tasks as an evaluation metric. Some researchers applied this indicator to measure the quality of the synthesised images during the harmonisation [125,126].

### 5.4. Main task based performance evaluation

Many studies demonstrated its effectiveness by comparing the performance of the main tasks before and after harmonisation methods. Although it is a result-orientated evaluation method and may be affected by the random initialisation of the main-task models (machine learning models), it can prove the effectiveness of the harmonisation method to some extent. The main tasks involved in harmonisation evaluation mainly include regression [57,68,91], segmentation, and classification [84]. The assessment is usually done by the Dice coefficient (Dice) or the intersection over union (IoU) [62,100,106,108,126,129,144]. On the contrary, the classification tasks are variously evaluated, e.g., using area under the receiver operating characteristics curve (AUC) [50,54,86,92, 106, 111, 115,138,143,146,154,159], accuracy [38,48,50,59, 60, 62, 63,69,77,89,99,106,131 133,140,143,145,159], true positive rate [67, 135], sensitivity [48,143], specificity [48] and Matthews correlation coefficient (MCC) [69]. Note that MCC is a balanced measurement for the binary classification tasks, with comprehensive evaluations of TP, TN, FP, and FN, therefore it is divided into the main task based performance evaluation.

### 5.5. Cohort classification

Different from comparing the variety of data or data-derived features before and after harmonisation, some studies reported the effectiveness of harmonisation strategies through adopting cohort classification [49, 63,94]. The core idea of this metric is that the cohort should be more difficult to identify when an effective harmonisation strategy is employed. For instance, Wachinger et al. [63] compared the accuracy of cohort identification of the raw dataset, dataset applied with z-score normalisation, linear model and ComBat, respectively. The worst results were gained by ComBat, which indicates the best harmonisation ability since the classifier cannot well identify the cohorts after harmonisation.

### 5.6. Visualisation

Visualisation refers to the techniques that can picture the data distribution from low dimensional feature space or sample intensities. Approaches for harmonisation assessment that visualising the data distribution in latent space including principal component analysis (PCA) [56,57,59,68,70,81,83,146], uniform manifold approximation and projection (UMAP) [38,50,77,133,135,142,160], and t-distributed stochastic neighbour embedding (t-SNE) [54,74,75,79, 83, 121,132,133, 143,145]. These approaches convert different high-dimensional data into low-dimensional data and plot them into the same feature space. The harmonisation is well performed if the visualisation of data distribution is mixed instead of assembling as different clusters. In addition to visualising data distributions, some researchers also plot the intensities or location-and-scales (mean and variance) of each sample before and after harmonisation to assess the performance [46,83,86,97, 98, 100, 103,106,107,125].

## 6. Applications of computational data harmonisation

Data harmonisation has been widely adopted in various fields of digital healthcare, including the manual harmonisation of tabular data, computational data harmonisation of the gene, radiomics and pathological data. Though there have been many efforts on removing the artefacts of time series signal data (such as EEG, ECG) [161], these works mainly focus on the removal of noises caused by biological variances. For instance, researchers employed filtering [162] and wavelet transform [163] approaches to remove the ocular, muscle and cardiac artefacts. However, these studies barely pay attention to the device/site variances, indicating the lack of harmonisation studies to these time series signal data. This section illustrates the application of computational data harmonisations in gene expression, radiomics and pathology.

### 6.1. Gene expression analysis

The process of generating a functional gene product from the information within a gene is referred to as gene expression, which is one of the major research areas in biomedical research. The traditional

approach of gene expression analysis is microarray technology, which relies on comprehensive chemical reactions to convert RNA to cDNA. The latest gene expression analysis method is single-cell RNA sequencing (scRNA-Seq). It isolates the single cells and RNA for transcription, library generation and sequencing, using the new generation sequencing (NGS) techniques. Unfortunately, due to the various NGS platforms and experimental environments (pH, temperature), both microarray technology and scRNA-sequencing are highly affected by cohort bias. Therefore, computational data harmonisation is widely used in microarray [46,71,72,80,82,84,85,89,93] and scRNA-Seq [56,67,68, 70,73 75,77,81,83, 121, 132, 133,135,140,142,145,160] to remove the cohort bias.

### 6.2. Radiomics analysis

Radiomics refers to the extraction and analysis of a large number of quantitative image features from medical images obtained by CT, PET, MRI, and other imaging modalities [164]. In addition to the studies on phantoms [61,65,94], research refer to MRI mainly focus on the brain [17,47 49,51 53,55,57,58, 63,66,86,87,90 92,95,97,99,100,115,117-119,125,126,129 131,143,144,157], breast [101,111], while those refer to CT focus on ear [60], liver [50,64] and lung [34,120]. In addition to CT, PET and MRI, optical coherence tomography also suffers severe scanner variability, which can be eased by computational data harmonisation approaches [129].

### 6.3. Pathology analysis

Most harmonisation strategies (also named stain/colour normalisation) in pathology aimed to address the stain variance. These studies mainly focused on uterus [108], breast [39,62,102,103,107,110 113], lymphoma [103,138], skin [105], liver [106], renal [59], and stomach [109].

### 6.4. Other modalities

Metabolomics is an omics technology that monitors and discovers metabolic changes in people in relation to illness state or in reaction to a medical or external intervention using current analytical instruments and pattern recognition algorithms. The nonlinear cohort bias during the liquid chromatography−mass spectrometry can be removed by computational methods [146].

## 7. Meta-analysis

In this review, the methodologies and metrics were grouped based on different ideas or theories, and the meta-analysis was conducted and reported in three areas/modalities. The reason why the results were explored and discussed through different modalities (gene, radiomics, and pathology) is that these data have different properties. For instance, data of gene analysis are expression matrices, data of radiomics are high-dimensional volume array (grayscale image per slice), and data of pathology are colour images with huge sizes.

### 7.1. Meta-analysis

**Data properties and study trends.** The number of studies and data properties for harmonisation approaches from 2000 to 2021 is demonstrated in the top left in Fig. 8, with the percentage of studies that was conducted on the public dataset. The public data can be acquired through open source websites or archives while the in-house data cannot be acquired. There has been a dramatic increase in the number of harmonisation studies since 2019, indicating an urgent need to conduct large-scale studies and data harmonisation strategies. In addition, we demonstrate the number of harmonisation studies on different sub-modalities in recent years (Microarray and sRNA-seq for gene

expression studies, CT and MRI for radiomic studies, and Pathology). In terms of gene expression, the harmonisation approaches for microarrays were mainly presented before 2015, while that for sRNA-seq have become the latest topic in recent years. As for radiomics studies, researchers have realised the importance to improve the reproducibility of radiomics features, especially in MRI. Data harmonisation for digital pathology has been noted in decades, while it receives more attention in recent years.

**Strategies and modalities.** Due to the diversity of biomedical data modalities, the relationship of different strategies and modalities was explored. As shown in Fig. 9, the distribution based methods were commonly applied in gene expression and radiomics studies, which account for 79% and 59% of the employed approaches, respectively; however, only a few of them (5%) were employed in digital pathology. The empirical Bayes methods dominated the distribution based methods because of their generalisation ability and robustness, while the RUV and SH were more commonly used in specific fields. The image processing approaches were mainly used in digital pathology, employing standardisation/normalisation and stain separation ideas to merge the multi-cohort data. Unlike the distribution based and image processing based methods, invariant feature learning and synthesis were found to be applicable in all three modalities recently, dominated by deep learning based algorithms.

**Evaluation metric.** The evaluation metric is another crucial aspect when developing computational data harmonisation strategies. It describes the performance of harmonisation methods via analysing the distribution, correlation and values between the source and target cohorts. amongst all evaluation metrics, visualisation was the most commonly used method to present harmonisation effects, followed by evaluating the main tasks (Fig. 10). Some studies tried to evaluate via classifying the cohorts, but this may have a limitation since the inability to distinguish cohorts does not mean that all data is well harmonised. Overall, even there are many options for the assessment of harmonisation strategies, there still exist some barriers to implementing harmonisation assessment in clinical flows. The evaluation can only be acquired when (1) there are paired datasets (which is inapplicable in real clinical settings); (2) there is a certain machine learning-based module for performance comparison (demands for well-trained computational modules); (3) there are clinicians for visual assessment (subjective and time-consuming); (4) there are predefined regions of interest (demands for manual annotation or computational modules). Moreover, data harmonization is of utmost importance for a seamless federation of models (i.e. a naïve federation approach in which no additional algorithms are needed to cope with incoherencies in local datasets). Therefore, to which extent should local datasets be harmonized not to destroy locally contextual particularities of the data that positively contribute to the local generalization of the models.

In addition to reporting the utilisation of different metrics, details of evaluation metrics in terms of different modalities were presented in Fig. 10. Visualisation is preferred for gene expression and digital pathology, including the visualisation of data distribution (e.g., using UMAP, t-SNE) and samples before and after harmonisation. Many pathology studies applied distribution based metrics, such as structure similarity and normalised median intensity. The task based evaluation was also considered reliable, which accounts for 19%, 28% and 13% of all evaluation metrics.

**Data scale-images.** The scale of samples in radiomics and pathology is closely related to image resolution and qualities. We analysed the variable $\varepsilon = \sqrt{w * h}$ of studies that involved images with width $w$ and height $h$. Most radiomics studies were performed with 256 pixels while some were conducted with 512 and 128 pixels, this was because of the shortage of GPU or RAMs especially when dealing with 3D or multi-slice datasets. Moreover, it is of note that 69.7% of radiomics studies did not report the image size, while that proportion of pathological studies was 20%. For pathological images, most studies were conducted with large

scale images ($\varepsilon > 500$), since the image processing algorithm performs better results when considering the larger field of views.

**Data scale-cohort.** In addition to the sample (image) scales, the scale of cohorts is also important for data harmonisation approaches. For gene expression, more than half of the studies were performed on large datasets (number of cases per cohort > 5000). Most radiomic studies were conducted on small datasets, with the number of cases (scans) per cohort < 200.

### 7.2. Research directions

This section first presents research tracks for different kinds of data harmonisation approaches based on their limitations, respectively, and states the common restrictions of previous studies.

**Directions for distribution based methods.** The distribution based methods mainly map the source data to the target one through the estimation of the cohort variances. This leads to issues that (1) most distribution based methods were conducted based on refined feature vectors that required prior knowledge of the region of interest. This prior knowledge conflicts with the original purpose that harmonisation approaches are proposed to process multicentre datasets to build robustness and precise computational tools because the region of interest cannot be well predicted by models trained without data harmonisation; (2) although studies have proved that some distribution based methods (such as ComBat) can remove cohort bias while preserving the differences between radiomics features on phantoms, all these methods cannot be well performed to images or high dimensional signals, due to demanding computational complexity; (3) the data harmonisation needs to be performed to entire datasets again when new data are added; (4) some approaches are pairwise (e.g., XPN, DWD, CCA, MNN, Seurat), leading to a complex training procedure (repeated training) when they are implemented to multicentre datasets (number of cohorts > 2). In particular, the first cohort will be considered as the target cohort to correct samples in the second one, and these corrected samples are then added to the first cohort [140].

To overcome these problems, researchers may (1) focus on the harmonisation of raw datasets, instead of data-derived features; (2) develop highly efficient data harmonisation approaches that can deal with a large amount of data; (3) enhance the robustness of data harmonisation strategies; (4) develop methods that can simultaneously harmonise multicentre datasets; and (5) avoid using pair-wise samples for algorithms development.

**Directions for image processing based methods.** Image processing based methods can harmonise the image data without complex procedures. However, these methods also have limitations as (1) some of them (such as stain separation) can be only performed to specific fields; (2) some (image filtering) methods heavily rely on empirical settings, such as filtering kernel sizes and kernel types, which are less efficient and hard to reproduce; and (3) some may lose the information during non-linear transforms. To address these issues, researchers should pay attention to general data harmonisation approaches that do not heavily rely on empirical settings.

**Directions for synthesis.** Although deep learning based synthesis solutions have advanced rapidly and achieved significant performance, these methods still suffer from poor reproducibility and generalisability. The obvious limitations are (1) most synthesis methods were built based on the existing multicentre datasets, which lack evaluations on new datasets; (2) the GAN based models are prone to instability and may hallucinate or introduce unrealistic changes; and (3) training a GAN based model requires a large amount of training data for all cohorts, which may be not feasible for clinical studies. To overcome these drawbacks, researchers should (1) report the data harmonisation performance on new datasets that are not involved during model development; (2) enhance the stability of data synthesis; (3) build data harmonisation strategies using less training data.

**Directions for invariant feature learning.** Invariant feature

**Table 4**
Data scale (image size) in previous studies.

|  | Small | Middle | Large | N/A |
|---|---|---|---|---|
| Radiomics | 9.1% (6) | 18.2% (12) | 3.0% (2) | 69.7% (46) |
| Pathology | 5.0% (1) | 15.0% (3) | 60.0% (12) | 20% (4) |

* The small, middle, large image sizes are defined as $\in [0, 256), \in [256, 512), \varepsilon > [512, \infty)$, respectively, N/A indicates there is no report of image size.

learning can reduce the disadvantages of synthesis approaches by learning how to extract cohort-invariant features from datasets, but it still faces some challenges. For instance, it can only extract invariant features for analysis while cannot obtain harmonised data. Therefore, future studies should focus on how to generate the harmonised data using extracted invariant features.

**Explainable AI and harmonisation studies.** Another research niche that still remains uncharted in the literature related to data harmonisation is the use of explainable Artificial Intelligence (XAI) methods [165] to identify possible reasons for incoherent data representations. We envision that XAI approaches can be exploited to gain insight on which visual artefacts are present in data instances that imprint a bias on the predicted outcome of a data-based model. This insight can be then analysed to decide whether the rooting cause of such biasing artefacts correspond to insufficient harmonisation of medical data before the learning phase. Furthermore, out of distribution examples can be also detected by virtue of local explanatory techniques (e.g., those capable of discerning which parts of the input to a model are pushing their output towards one class or another), which upon inspection can be attributed to other exogenous phenomena that can relate to data harmonisation, such as a possible miscalibration of the medical equipment or a change in the protocols capturing the data themselves. On the other hand, better harmonisation has benefits to XAI, since all the data are harmonised into the same standard and no cohort biases would be introduced to the XAI system [166,167]. All in all, we foresee an interesting research cross-fertilization at the crossroads between harmonisation and XAI.

**Limitations for methodology design.** Most studies for data harmonisation did not follow a stepwise design methodology, which cannot be reproduced easily by third parties. For instance, as shown in Table 4, more than half of the radiomic studies did not report the image scale. Moreover, the different definitions of 'reproducible' in previous studies and various evaluation metrics greatly hinder the method comparison for further research.

## 8. Checklist and guidance

To address the issues of methodology design, we presented a Checklist for Computational Data Harmonisation in Digital Healthcare (CHECDHA) to enhance the reproducibility and methodological principle, inspired by the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [169]. Furthermore, the guidance on how to choose data harmonisation strategies is also presented in this section.

### 8.1. Checklist criteria

The proposed checklist clarifies the common practice for data harmonisation through data, model, evaluation, result, and discussion, shown in Table 5.

The proposed CHECDHA checklist can greatly standardise the process of data harmonisation studies, which comprehensively describe the motivation, data, data harmonisation strategy, evaluation and conclusions. Start with a clear motivation (Fig. 12), researchers should first emphasise the importance of performing data harmonisation in a certain field. Then, the compositions of datasets should be illustrated in detail, including the common and specific attributes shown in the checklist. When introducing methodologies, the authors should clearly state their ideas and implementation details (input domain, architecture, input

**Table 5**

Checklist for Computational Data Harmonisation in Digital Healthcare (CHECDHA) criteria.

| Category | | Item | Explanation | Example |
|---|---|---|---|---|
| Motivation | | Background | The application field of the dataset(s) | Information fusion of DW-MRI data from different scanners |
| | | Importance | Why this study is conducted, how important it is | Dramatically increase the statistical power and sensitivity of clinical studies |
| Data | Common | Dataset | What the dataset(s) is (are), how it is (they are) collected (details of acquisition protocols, entry and exit criteria) How many categories, cohorts, subjects, and cases are included in the studies | $m$ healthy subjects under $n$ protocols ($m \times n$ cases, $n$ cohorts) Protocol 1: … Protocol 2: … |
| | | Property | Whether the dataset (s) is (are) in-house or public, provide the access link if appropriate | Public/In-house |
| | | Pre-processing | How the dataset is pre-processed | Z-score normalisation |
| | | Ground truth | What the ground truth is and how it is generated | Cohort $x$ under protocol $i$ |
| | | Partition | For machine learning, how the dataset is partitioned into training, validation, and testing subsets in terms of the number of samples, patients | 7:2:1 for training, validation and test |
| | | Augmentation | For machine learning, how the dataset is augmented | Randomized flip, rotation |
| | Specific | MRI sequence | What the MRI sequence is | Diffusion-weighted |
| | | Region | Which region(s) of the body or the subject in the dataset is (are) covered | Brain |
| | | Slice size | What the sizes of each slice are | $512 \times 512$ |
| | | Pixel/Voxel size | What the physical length of a pixel/voxel is | 0.25 mm/ $1mm^3$ |
| | | WSI size | What the sizes of the whole slide images are | $12,000 \times 30,000$ |
| | | Patch size | What the extracted image patches are | $256 \times 256$ |
| | | mmp | What the microns per pixel in the level-0 scan are | – |
| Model | | Workflow | What the procedures of train and inference are, illustrated by the flow chart(s) if appropriate. | – |
| | | Learning approaches | What the learning method is. e.g., supervised learning, un/semi-supervised learning | Semi-supervised learning |
| | | Architecture | What the structure of the proposed neural network is, if appropriate | nnUNet |

**Table 5** (*continued*)

| | Task | The description of main tasks conducted on harmonised datasets, e.g., lesion segmentation/classification. | Tumour Segmentation |
|---|---|---|---|
| | Input domain | What the input modality of the proposed method is | 3-D images / 2D feature vectors |
| | Input size | The input sizes of the model | $n \times w \times h \times c$ |
| | Loss | What the optimisation functions are during the training. | Dice and cross-entropy loss |
| | Open-source | Whether the source code is available or not, provide the link if appropriate. | Open-source code www.github.com... |
| | Platform | The learning library used to build the model | TensorFlow 2.5.0 |
| Evaluation | Statistical Analysis | What the evaluation methods of statistical analysis are | ANOVA-test |
| | Metric | What indicators are used to evaluate harmonisation performance, e.g., the ratio of the reproducible features, coefficient of variation, Pearson correlation coefficient. | Intra-class correlation coefficient ($>0.9$ is considered reproducible) |
| | Comparison | What existing approaches are used to compare the performance of the proposed method | stVAE |
| | Visualisation | What approaches are used to visualise the data distribution before and after harmonisation strategies | t-SNE/UMAP/PCA |
| Result | Result | What the quantitative values of evaluation metrics are. | – |
| | Time-consuming | The computational time of the proposed method and the comparisons. | 30 s per case |
| Discussion | Novelty | What the innovation of the proposed method is. | – |
| | Strength | The importance/significance of the issue addressed by the proposed method. | – |
| | Limitation | What remained and unsolved issues are. | – |
| | Future works | Whether there will be potential studies in the future. | – |

size, development platform, etc.). During the evaluation, researchers should assess the reproducibility using new/independent data or data-derived features before and after data harmonisation by appropriate metrics. Meanwhile, the data harmonisation performance of previous approaches should be considered as comparisons to reflect the advantages of the proposed method. At last, the novelty, strength, limitations and future works should be given in the discussion and conclusion sections.
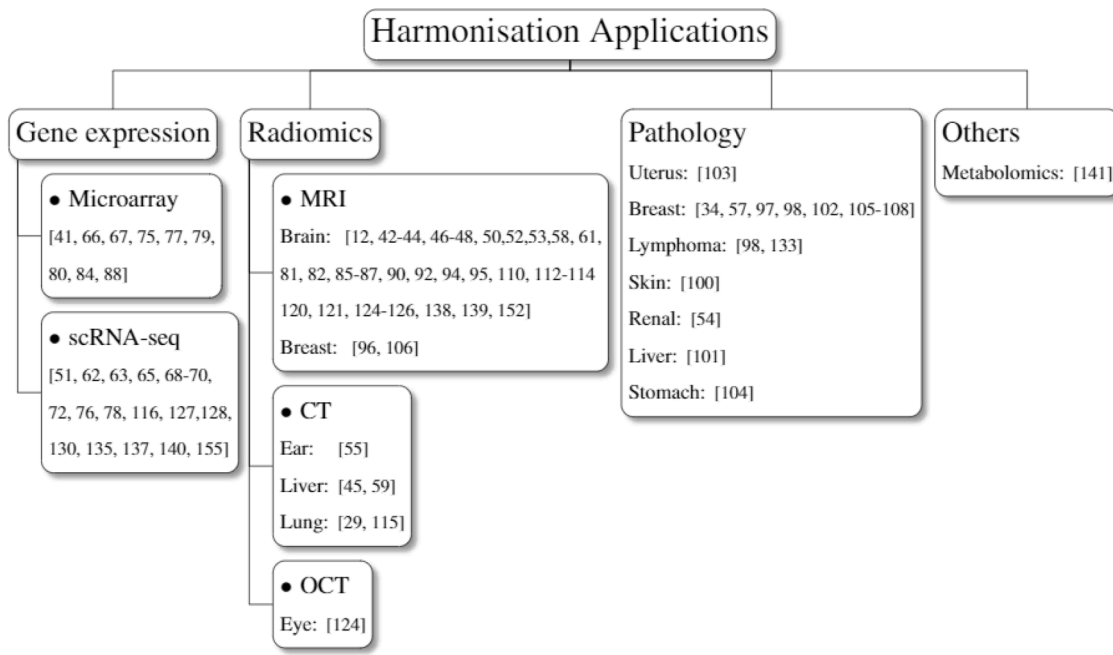
**Fig. 7.** Taxonomy of applications that involved computational data harmonisation strategies.

### 8.2. Guidance of data harmonisation strategies and metrics

Studies have shown that implementing inaccurate data harmonisation strategies may lead to significant bias, which results in more inaccurate predictions [168]. To guide the method selection, a flowchart presenting possible ways of data harmonisation is presented in Fig. 13. As the flowchart illustrates, the distribution based methods can be well performed on refined features or gene matrices. For high dimensional images, image processing methods are recommended when a high-performance GPU is not available. The deep learning based methods (including invariant feature learning and synthesis) can be applied to all kinds of modalities, while it requires sufficient training samples. The invariant feature learning methods are recommended when the main task can be integrated with the training process, since the synthesis may introduce unrealistic artefacts to the data.

For evaluation, the selection of metrics can directly affect whether
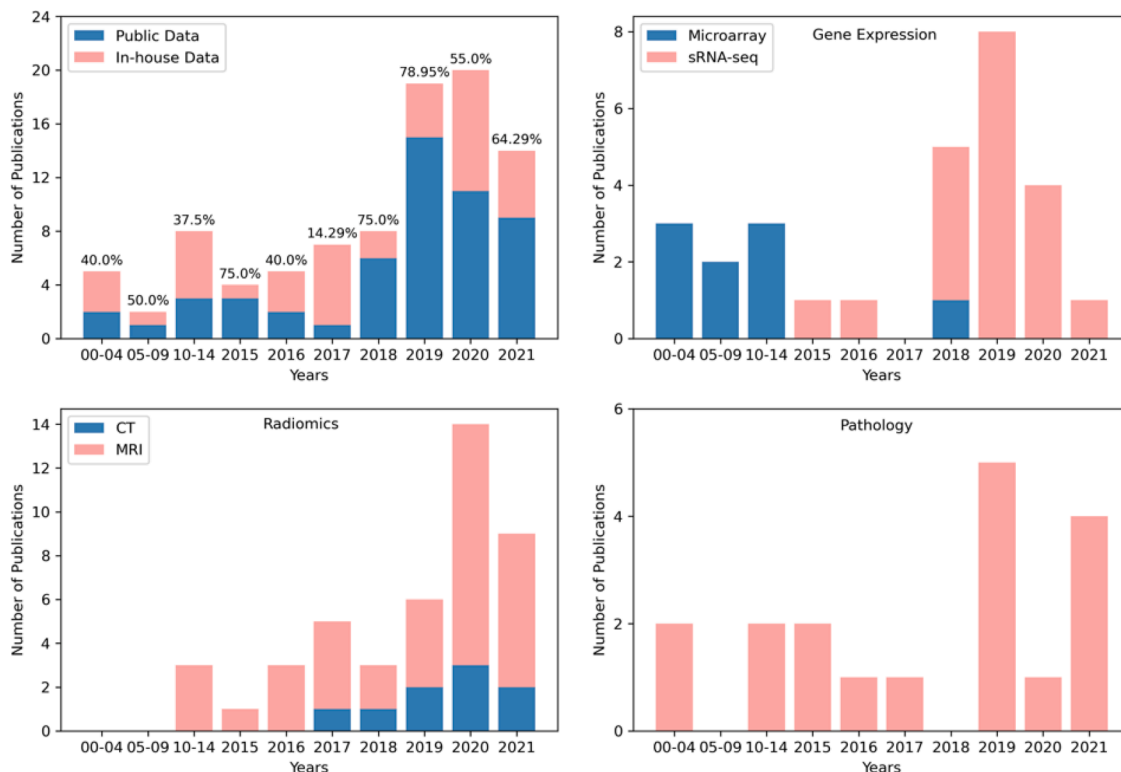


**Fig. 8.** Number of publications and years in terms of data properties and modalities. The public data is the open source data that can be acquired, the in-house data is not available from the internet. The percentage in the top left subfigure is the ratio of studies that were conducted on the public dataset.
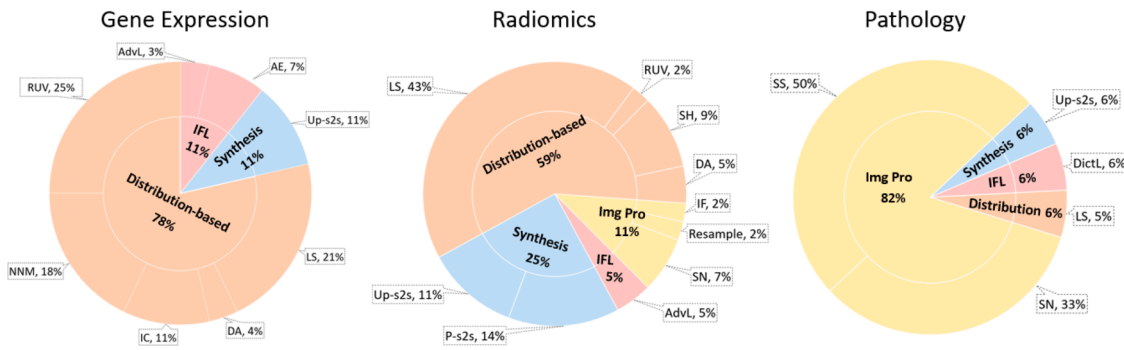
**Fig. 9.** Harmonisation strategies in terms of different modalities. 'IFL' indicates invariant feature learning approaches, "Img Pro" refers to image processing approaches. The percentage of sub-methods is annotated with the abbreviations of sub-methods in each pie chart.

the results are reliable or not. Here we summarise and recommend data harmonisation metrics based on different conditions in Fig. 14. Visualisation is the most intuitional way to analyse data harmonisation results, which can be implemented by visualising the raw data with t-SNE/UMAP/PCA or visualising the data harmonised raw data. Main task based evaluation can directly illustrate the effectiveness of the data harmonisation strategies, by comparing the main task performance on data before and after the data harmonisation. If the harmonised ground truth is not available, one can use distribution based metrics to assess the degree of sample mixture (although this may require the cohort label).

When the harmonised ground truth can be acquired, the value based or correlation based metrics can precisely present the data harmonisation performance.

## 9. Conclusion

Computational data harmonisation has been proposed for digital healthcare research studies in decades. However, bridging basic science research models and data fusion into multicentre, multimodal and multi-scanner medical practice and clinical trials can be challenging unless
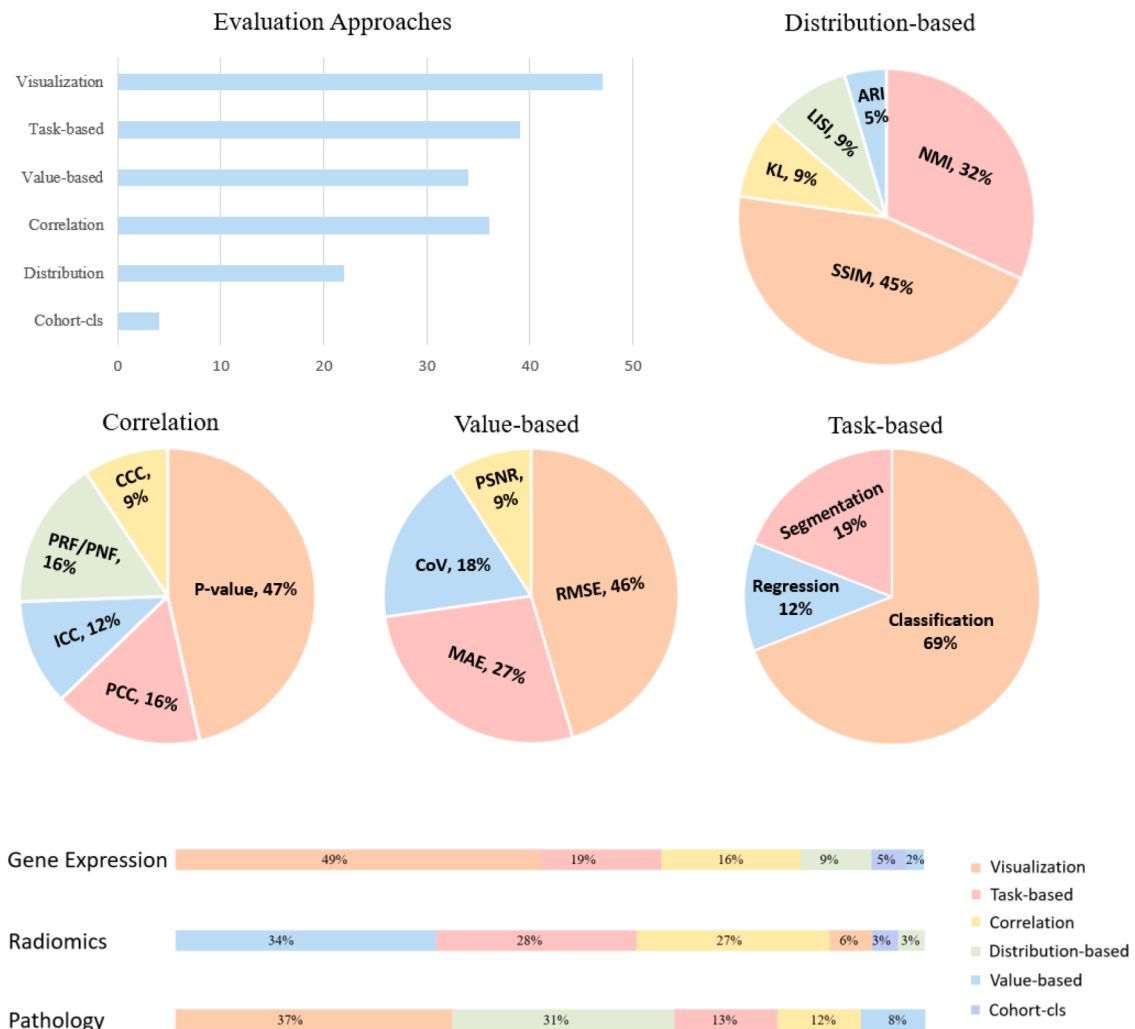


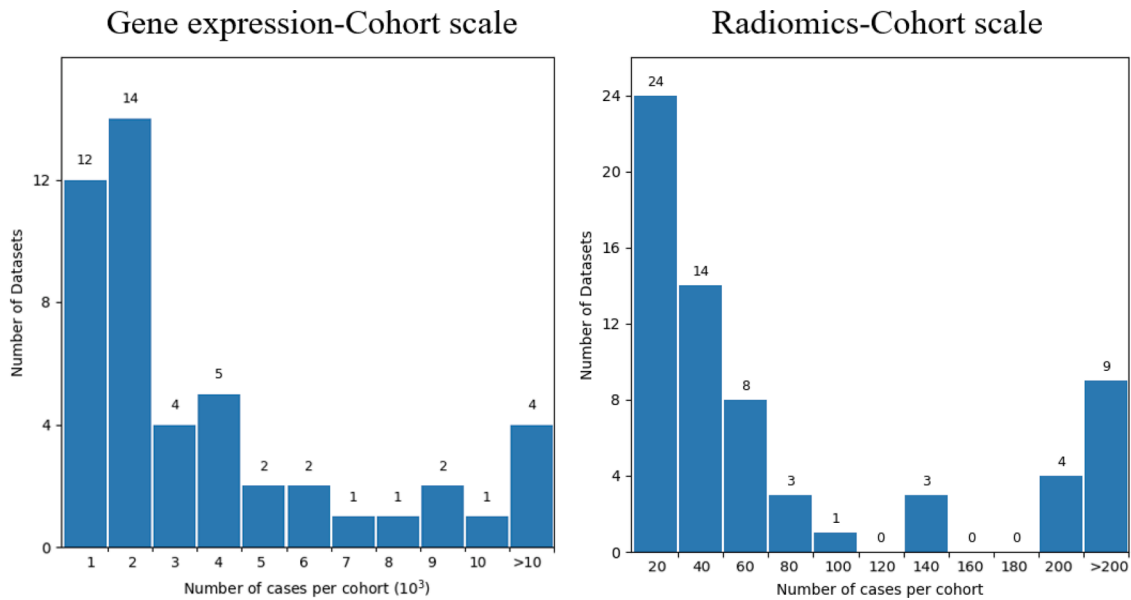**Fig. 10.** Evaluation metrics in terms of different modalities.

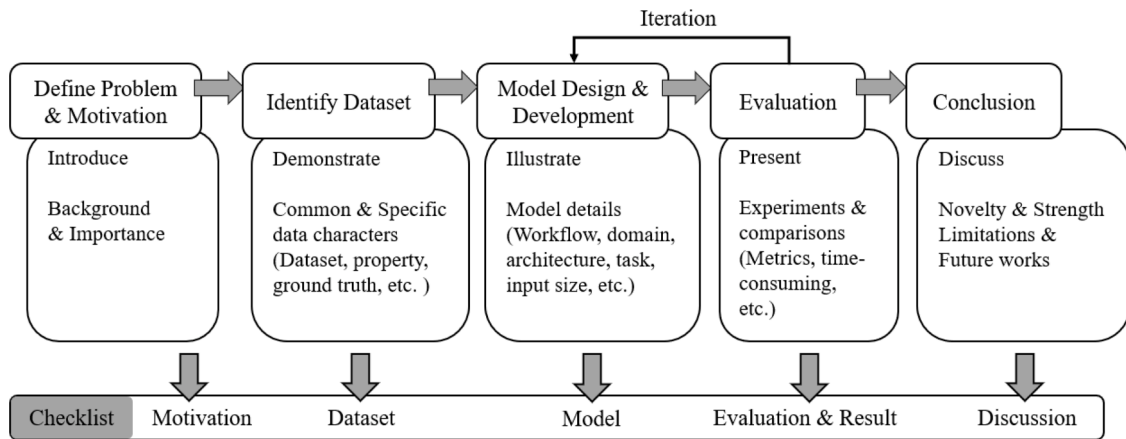**Fig. 11.** Scales of cohorts in gene expression and radiomics studies.



**Fig. 12.** Workflow of conducting data harmonisation studies guided by the checklist.

data harmonisation can be performed effectively. Furthermore, transfer/federated/multitask learning and other areas wherein knowledge is exchanged amongst models only work under ideal conditions, whenever the distribution shift is not large enough for the exchange knowledge to remain coherent across models/centres working over different data sources. Otherwise, data harmonisation is needed. Unfortunately, it is unclear which approaches and metrics should be employed when dealing with multimodal datasets. Moreover, there lacks a 'standardised' stepwise design methodology, which leads to poor reproducibility of the existing studies.

To overcome these issues, this paper summarises and categorises the existing data harmonisation strategies and metrics based on different theories, and subsequently presents the CHECDHA criteria. The proposed CHECDHA criteria help researchers to conduct data harmonisation studies in a standardised format, which can greatly advance academic reproducibility and development. Moreover, data harmonisation approaches and evaluation metrics in terms of three modalities are summarised to help researchers to select appropriate strategies (Fig. 7 and Fig. 8). In addition to summarising the methodologies, guidance of method and metrics selection (Fig. 11 and Fig. 12) is also provided according to the different conditions. Last but not least, limitations and directions of different methods are illustrated for future
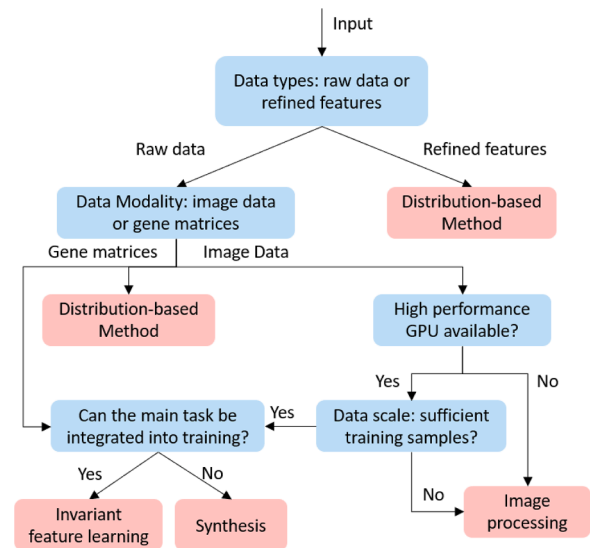


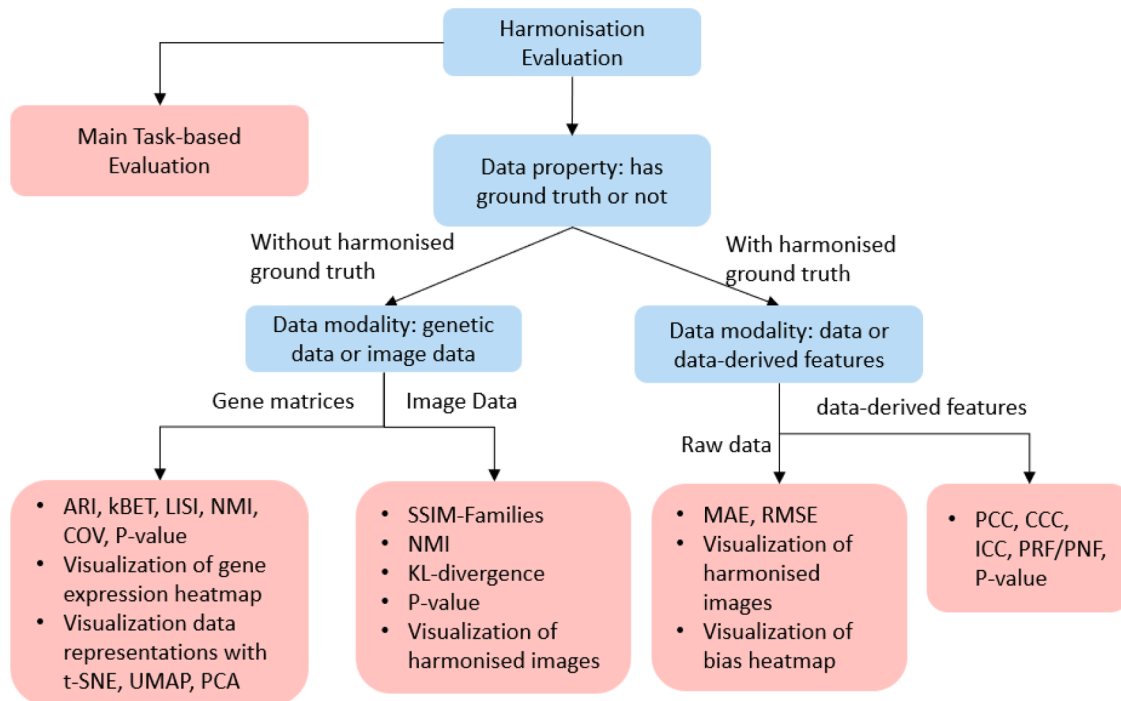**Fig. 13.** Flowchart of how to select data harmonisation strategies.

**Fig. 14.** Flowchart of how to select harmonisation metrics.

works.

Data harmonisation, an important process in large multicentre studies, has drawn more and more attention in computational biomedical research. It can be well adapted to a federated learning system to promote the development of computational modules and plays an important role in biomedical research including radiomic, genetic and pathological studies. Due to the lack of criteria when reporting research findings of harmonisation studies, we strongly appeal that the researchers should follow and expand the checklist presented in this survey.

## Author statements

YN, JDS, FH and GY conceived and designed the study, contributed to data analysis, contributed to data interpretation, and contributed to the writing of the report. YN, JDS, SW1, CS, MR, IS, KH, JO, JN, JG, BE, AP, AAB, MIM, SW2, WV, NF, JPC, EVR, AC, HW, PL, LCA, LMB, FH, and GY contributed to the literature search. YN contributed to data collection and performed data curation and contributed to the tables and figures. YN, JDS, LCA, LMB, and GY contributed to Writing - Review & Editing. SW1, CS, KH, JG, AAB, MIM, SW2, WV, EVR, PL, LMB and GY contributed to Funding acquisition. NF contributed to Project administration. FH oversaw the study. GY supervised the work. All authors contributed to the article and approved the submitted version.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

[#] DRAGON Consortium:

Xiaodan Xing[a], Ming Li[a], Scott Wagers[b], Rebecca Baker[c], Cosimo Nardi[d], Brice van Eeckhout[e], Paul Skipp[f], Pippa Powell[g], Miles Carroll[h], Alessandro Ruggiero[i], Muhunthan Thillai[i], Judith Babar[i], Evis Sala[i], William Murch[j], Julian Hiscox[k], Diana Baralle[l], Nicola Sverzellati[m]

[##] CHAIMELEON Consortium:

Ana Miguel Blanco[n], Fuensanta Bellvís Bataller[o], Mario Aznar[p], Amelia Suarez[p], Sergio Figueiras[q], Katharina Krischak[r], Monika Hierath[r], Yisroel Mirsky[s], Yuval Elovici[s], Jean Paul Beregi[t], Laure Fournier[t], Francesco Sardanelli[u], Tobias Penzkofer[v], Karine Seymour[w], Nacho Blanquer[x], Emanuele Neri[y], Andrea Laghi[z], Manuela França[aa], Ricard Martinez[ab]

[a] National Heart and Lung Institute, Imperial College London, London, UK

[b] BioSci Consulting, Maasmechelen, Belgium

[c] Clinical Data Interchange Standards Consortium, Austin, Texas, United States

[d] University of Florence, Firenze, Italy

[e] Medical Cloud Company, Liège, Belgium

[f] TopMD, Southampton, UK

[g] European Lung Foundation, Sheffield, UK

[h] Department of Health, Public Health England, London, UK

[i] Department of Radiology, University of Cambridge, Cambridge, UK

[j] Owlstone Medical, Cambridge, UK

[k] University of Liverpool, Liverpool, UK

[l] University of Southampton, Southampton, UK

[m] University of Parma, Parma, Italy

[n] Medical Imaging Department, Hospital Universitari i Politècnic La Fe, Valencia, Spain

[o] QUIBIM, Valencia, Spain

[p] Matical Innovation, Madrid, Spain

ᑫ Bahía Software, A Coruña, Spain

ʳ European Institute for Biomedical Imaging Research, Vienna, Austria

ˢ Ben Gurion University of the Negev, Be'er Sheva, Israel

ᵗ Le Collège des Enseignants en Radiologie de France, France

ᵘ Research Hospital Policlinico San Donato, Milan, Italy

ᵛ Charité – Universitätsmedizin Berlin, Berlin, Germany

ʷ Medexprim, Labège, France

ˣ Valencia Polytechnic University, Valencia, Spain

ʸ University of Pisa, Pisa, Italy

ᶻ Sapienza University of Rome, Rome, Italy

ᵃᵃ The Centro Hospitalar Universitário do Porto, Portugal

ᵃᵇ University of Valencia, Valencia, Spain

## References

1 W.T. Clarke, O. Mougin, I.D. Driver, C. Rua, A.T. Morgan, M. Asghar, S. Clare, S. Francis, R.G. Wise, C.T. Rodgers, Multi-site harmonization of 7 tesla MRI neuroimaging protocols, Neuroimage 206 (2020), 116335.

2 D. Delbeke, R.E. Coleman, M.J. Guiberteau, M.L. Brown, H.D. Royal, B.A. Siegel, D. W. Townsend, L.L. Berland, J.A. Parker, K. Hubner, Procedure guideline for tumor imaging with 18F-FDG PET/CT 1.0, J. Nucl. Med. 47 (2006) 885–895.

3 J. Simon, D. Li, A. Traboulsee, P. Coyle, D. Arnold, F. Barkhof, J. Frank, R. Grossman, D. Paty, E. Radue, Standardized MR imaging protocol for multiple sclerosis: consortium of MS Centers consensus guidelines, Am. J. Neuroradiol. 27 (2006) 455–461.

4 B.-.M. Schmidt, C.J. Colvin, A. Hohlfeld, N. Leon, Defining and conceptualising data harmonisation: a scoping review protocol, Syst. Rev. 7 (2018) 1–6.

5 B.-.M. Schmidt, C.J. Colvin, A. Hohlfeld, N. Leon, Definitions, components and processes of data harmonisation in healthcare: a scoping review, BMC Med. Inform. Decis. Mak. 20 (2020) 1–19.

6 R. Da-Ano, D. Visvikis, M. Hatt, Harmonization strategies for multicenter radiomics investigations, Phy. Med. Biol. 65 (2020) 24TR02.

7 M.S. Pinto, R. Paolella, T. Billiet, P. Van Dyck, P.-.J. Guns, B. Jeurissen, A. Ribbens, A.J. den Dekker, J. Sijbers, Harmonization of brain diffusion MRI: concepts and methods, Front. Neurosci. 14 (2020) 396.

8 S. Gitto, R. Cuocolo, D. Albano, F. Morelli, L.C. Pescatori, C. Messina, M. Imbriaco, L.M. Sconfienza, CT and MRI radiomics of bone and soft-tissue sarcomas: a systematic review of reproducibility and validation strategies, Insights Imaging 12 (2021) 1–14.

9 S.A. Mali, A. Ibrahim, H.C. Woodruff, V. Andrearczyk, H. Müller, S. Primakov, Z. Salahuddin, A. Chatterjee, P. Lambin, Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods, J. Pers. Med. 11 (2021) 842.

10 T. Chen, M. Philip, K.-A.Lê Cao, S. Tyagi, A multi-modal data harmonisation approach for discovery of COVID-19 drug targets, Brief. Bioinform. (2021).

11 C.M. Tax, F. Grussu, E. Kaden, L. Ning, U. Rudrapatna, C.J. Evans, S. St-Jean, A. Leemans, S. Koppers, D. Merhof, Cross-scanner and cross-protocol diffusion MRI data harmonisation: a benchmark database and evaluation of algorithms, Neuroimage 195 (2019) 285–299.

12 D.M. Hutchinson, E. Silins, R.P. Mattick, G.C. Patton, D.M. Fergusson, R. Hayatbakhsh, J.W. Toumbourou, C.A. Olsson, J.M. Najman, E. Spry, How can data harmonisation benefit mental health research? An example of the Cannabis cohorts research consortium, Australian New Zealand J. Psychiat. 49 (2015) 317–323.

13 B. Zhao, Y. Tan, W.-.Y. Tsai, J. Qi, C. Xie, L. Lu, L.H. Schwartz, Reproducibility of radiomics for deciphering tumor phenotype with imaging, Sci. Rep. 6 (2016) 1–7.

14 N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, A. Sethi, A dataset and a technique for generalized nuclear segmentation for computational pathology, IEEE Trans. Med. Imaging 36 (2017) 1550–1560.

15 M. Hotta, R. Minamimoto, Y. Gohda, K. Miwa, K. Otani, T. Kiyomatsu, H. Yano, Prognostic value of 18 F-FDG PET/CT with texture analysis in patients with rectal cancer treated by surgery, Ann. Nucl. Med. (2021) 1–10.

16 M.V. Mattoli, M.L. Calcagni, S. Taralli, L. Indovina, B.S. Spottiswoode, A. Giordano, How often do we fail to classify the treatment response with [18 F] FDG PET/CT acquired on different scanners? Data from clinical oncological practice using an automatic tool for SUV harmonization, Mol. Imaging Biol. 21 (2019) 1210–1219.

17 H. Mirzaalian, L. Ning, P. Savadjiev, O. Pasternak, S. Bouix, O. Michailovich, G. Grant, C.E. Marx, R.A. Morey, L.A. Flashman, Inter-site and inter-scanner diffusion MRI data harmonization, Neuroimage 135 (2016) 311–323.

18 T. Zhu, R. Hu, X. Qiu, M. Taylor, Y. Tso, C. Yiannoutsos, B. Navia, S. Mori, S. Ekholm, G. Schifitto, Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study, Neuroimage 56 (2011) 1398–1411.

19 J. Jovicich, M. Marizzoni, B. Bosch, D. Bartrés-Faz, J. Arnold, J. Benninghoff, J. Wiltfang, L. Roccatagliata, A. Picco, F. Nobili, Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects, Neuroimage 101 (2014) 390–403.

20 P. Leo, G. Lee, N.N. Shih, R. Elliott, M.D. Feldman, A. Madabhushi, Evaluating stability of histomorphometric features across scanner and staining variations:

prostate cancer diagnosis from whole slide images, J. Med. Imaging 3 (2016), 047502.

21 R. Berenguer, M.D.R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M. V. Villas, F. Mansilla Legorburo, S. Sabater, Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters, Radiology 288 (2018) 407–415.

22 J.J. Sunderland, P.E. Christian, Quantitative PET/CT scanner performance characterization based upon the society of nuclear medicine and molecular imaging clinical trials network oncology clinical simulator phantom, J. Nucl. Med. 56 (2015) 145–152.

23 A. Jha, S. Mithun, V. Jaiswar, U. Sherkhane, N. Purandare, K. Prabhash, V. Rangarajan, A. Dekker, L. Wee, A. Traverso, Repeatability and reproducibility study of radiomic features on a phantom and human cohort, Sci. Rep. 11 (2021) 1–12.

24 N. Emaminejad, M.W. Wahi-Anwar, G.H.J. Kim, W. Hsu, M. Brown, M. McNitt-Gray, Reproducibility of lung nodule radiomic features: multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters, Med. Phys., (2021).

25 M. Kim, S.C. Jung, J.E. Park, S.Y. Park, H. Lee, K.M. Choi, Reproducibility of radiomic features in SENSE and compressed SENSE: impact of acceleration factors, Eur. Radiol. (2021) 1–14.

26 R. Yamashita, J. Perrin, J. Chakraborty, J.F. Chou, N. Horvat, M.A. Koszalka, A. Midya, M. Gonen, P. Allen, W.R. Jarnagin, Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation, Eur. Radiol. 30 (2020) 195–205.

27 S. Fiset, M.L. Welch, J. Weiss, M. Pintilie, J.L. Conway, M. Milosevic, A. Fyles, A. Traverso, D. Jaffray, U. Metser, Repeatability and reproducibility of MRI-based radiomic features in cervical cancer, Radiother. Oncol. 135 (2019) 107–114.

28 E. Saeedi, A. Dezhkam, J. Beigi, S. Rastegar, Z. Yousefi, L.A. Mehdipour, H. Abdollahi, K. Tanha, Radiomic feature robustness and reproducibility in quantitative bone radiography: a study on radiologic parameter changes, J. Clin. Densitom. 22 (2019) 203–213.

29 M. Meyer, J. Ronald, F. Vernuccio, R.C. Nelson, J.C. Ramirez-Giraldo, J. Solomon, B.N. Patel, E. Samei, D. Marin, Reproducibility of CT radiomic features within the same patient: influence of radiation dose and CT reconstruction settings, Radiology 293 (2019) 583–591.

30 T. Perrin, A. Midya, R. Yamashita, J. Chakraborty, T. Saidon, W.R. Jarnagin, M. Gonen, A.L. Simpson, R.K. Do, Short-term reproducibility of radiomic features in liver parenchyma and liver malignancies on contrast-enhanced CT imaging, Abdominal Radiol. 43 (2018) 3271–3278.

31 A. Midya, J. Chakraborty, M. Gönen, R.K. Do, A.L. Simpson, Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility, J. Med. Imaging 5 (2018), 011020.

32 B.A. Altazi, G.G. Zhang, D.C. Fernandez, M.E. Montejo, D. Hunt, J. Werner, M. C. Biagioli, E.G. Moros, Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms, J. Appl. Clin. Med. Phy. 18 (2017) 32–48.

33 P. Hu, J. Wang, H. Zhong, Z. Zhou, L. Shen, W. Hu, Z. Zhang, Reproducibility with repeat CT in radiomics study for rectal cancer, Oncotarget 7 (2016) 71440.

34 J. Choe, S.M. Lee, K.-.H. Do, G. Lee, J.-.G. Lee, S.M. Lee, J.B. Seo, Deep learning–based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses, Radiology 292 (2019) 365–373.

35 A.N. Primak, C.H. McCollough, M.R. Bruesewitz, J. Zhang, J.G. Fletcher, Relationship between noise, dose, and pitch in cardiac multi–detector row CT, Radiographics 26 (2006) 1785–1794.

36 D.S. Gierada, A.J. Bierhals, C.K. Choong, S.T. Bartel, J.H. Ritter, N.A. Das, C. Hong, T.K. Pilgram, K.T. Bae, B.R. Whiting, Effects of CT section thickness and reconstruction kernel on emphysema quantification: relationship to the magnitude of the CT emphysema index, Acad. Radiol. 17 (2010) 146–156.

37 P.-.Y. Tung, J.D. Blischak, C.J. Hsiao, D.A. Knowles, J.E. Burnett, J.K. Pritchard, Y. Gilad, Batch effects and the effective design of single-cell gene expression studies, Sci. Rep. 7 (2017) 1–15.

38 T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W.M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data, Cell 177 (2019) 1888–1902, e1821.

39 S. Vijh, M. Saraswat, S. Kumar, A new complete color normalization method for H&E stained histopatholgical images, Appl. Intell. (2021) 1–14.

40 D.E. Chandler, R.W. Roberson, Bioimaging: current concepts in light and electron microscopy, (2009).

41 D. Sun, J. Wang, Y. Han, X. Dong, J. Ge, R. Zheng, X. Shi, B. Wang, Z. Li, P. Ren, TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment, Nucleic Acids Res. 49 (2021) D1420–D1430.

42 M. Shafiq-ul-Hassan, G.G. Zhang, K. Latifi, G. Ullah, D.C. Hunt, Y. Balagurunathan, M.A. Abdalah, M.B. Schabath, D.G. Goldgof, D. Mackin, Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels, Med. Phys. 44 (2017) 1050–1062.

43 D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A.K. Jones, L. Court, Measuring CT scanner variability of radiomics features, Invest. Radiol. 50 (2015) 757.

44 G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M.G. Kramberger, The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study, Med. Image Anal. 66 (2020), 101714.

45 S. Rathore, S. Bakas, H. Akbari, G. Shukla, M. Rozycki, C. Davatzikos, Deriving stable multi-parametric MRI radiomic signatures in the presence of inter-scanner

variations: survival prediction of glioblastoma via imaging pattern analysis and machine learning techniques, in: Proceedings of the Medical Imaging 2018: Computer-Aided Diagnosis, International Society for Optics and Photonics, 2018, 1057509.

46 W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (2007) 118–127.

47 U. Pandey, J. Saini, M. Kumar, R. Gupta, M. Ingalhalikar, Normative baseline for radiomics in Brain MRI: evaluating the robustness, regional variations, and reproducibility on FLAIR Images, J. Magn. Reson. Imaging (2020).

48 M. Ingalhalikar, S. Shinde, A. Karmarkar, A. Rajan, D. Rangaprakash, G. Deshpande, Functional connectivity-based prediction of Autism on site harmonized ABIDE dataset, IEEE Trans. Biomed. Eng. (2021).

49 K. Wengler, C. Cassidy, M. van Der Pluijm, J.J. Weinstein, A. Abi-Dargham, E. van de Giessen, G. Horga, Cross-scanner harmonization of neuromelanin-sensitive MRI for multisite studies, J. Magn. Reson. Imaging (2021).

50 H. Beaumont, A. Iannessi, A.-.S. Bertrand, J.M. Cucchi, O. Lucidarme, Harmonization of radiomic feature distributions: impact on classification of hepatic tissue in CT imaging, Eur. Radiol. (2021) 1–10.

51 R. Garcia-Dias, C. Scarpazza, L. Baecker, S. Vieira, W.H. Pinaya, A. Corvin, A. Redolfi, B. Nelson, B. Crespo-Facorro, C. McDonald, Neuroharmony: a new tool for harmonizing volumetric MRI data from unseen scanners, Neuroimage 220 (2020).

52 J.C. Beer, N.J. Tustison, P.A. Cook, C. Davatzikos, Y.I. Sheline, R.T. Shinohara, K. A. Linn, A.s.D.N. Initiative, Longitudinal combat: a method for harmonizing longitudinal multi-scanner imaging data, Neuroimage 220 (2020), 117129.

53 J. Radua, E. Vieta, R. Shinohara, P. Kochunov, Y. Quidé, M.J. Green, C.S. Weickert, T. Weickert, J. Bruggemann, T. Kircher, Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA, Neuroimage 218 (2020), 116956.

54 H.M. Whitney, H. Li, Y. Ji, P. Liu, M.L. Giger, Harmonization of radiomic features of breast lesions across international DCE-MRI datasets, J. Med. Imaging 7 (2020), 012707.

55 M. Yu, K.A. Linn, P.A. Cook, M.L. Phillips, M. McInnis, M. Fava, M.H. Trivedi, M. M. Weissman, R.T. Shinohara, Y.I. Sheline, Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data, Hum Brain Mapp. 39 (2018) 4213–4227.

56 A. Espín-Pérez, C. Portier, M. Chadeau-Hyam, K. van Veldhoven, J.C. Kleinjans, T. M. de Kok, Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data, PLoS ONE 13 (2018), e0202947.

57 J.-.P. Fortin, N. Cullen, Y.I. Sheline, W.D. Taylor, I. Aselcioglu, P.A. Cook, P. Adams, C. Cooper, M. Fava, P.J. McGrath, Harmonization of cortical thickness measurements across scanners and sites, Neuroimage 167 (2018) 104–120.

58 J.-.P. Fortin, D. Parker, B. Tunç, T. Watanabe, M.A. Elliott, K. Ruparel, D.R. Roalf, T. D. Satterthwaite, R.C. Gur, R.E. Gur, Harmonization of multi-site diffusion tensor imaging data, Neuroimage 161 (2017) 149–170.

59 S. Kothari, J.H. Phan, T.H. Stokes, A.O. Osunkoya, A.N. Young, M.D. Wang, Removing batch effects from histopathological images for enhanced cancer diagnosis, IEEE J. Biomed. Health Inform. 18 (2013) 765–772.

60 C.T. Arendt, D. Leithner, M.E. Mayerhoefer, P. Gibbs, C. Czerny, C. Arnoldner, I. Burck, M. Leinung, Y. Tanyildizi, L. Lenga, Radiomics of high-resolution computed tomography for the differentiation between cholesteatoma and middle ear inflammation: effects of post-reconstruction methods in a dual-center study, Eur. Radiol. 31 (2021) 4071–4078.

61 A. Ibrahim, T. Refaee, R.T. Leijenaar, S. Primakov, R. Hustinx, F.M. Mottaghy, H. C. Woodruff, A.D. Maidment, P. Lambin, The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset, PLoS ONE 16 (2021), e0251147.

62 J. Lan, S. Cai, Y. Xue, Q. Gao, M. Du, H. Zhang, Z. Wu, Y. Deng, Y. Huang, T. Tong, Unpaired stain style transfer using invertible neural networks based on channel attention and long-range residual, IEEE Access 9 (2021) 11282–11295.

63 C. Wachinger, A. Rieckmann, S. Pölsterl, A.s.D.N. Initiative, Detect and correct bias in multi-site neuroimaging datasets, Med. Image Anal. 67 (2021), 101879.

64 J.J. Foy, H.A. Al-Hallaq, V. Grekoski, T. Tran, K. Guruvadoo, S.G. Armato Iii, W. F. Sensakovic, Harmonization of radiomic feature variability resulting from differences in CT image acquisition and reconstruction: assessment in a cadaveric liver, Phy. Med. Biol. 65 (2020), 205008.

65 M.-J.Saint Martin, F. Orlhac, P. Akl, F. Khalid, C. Nioche, I. Buvat, C. Malhaire, F. Frouin, A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study, magnetic resonance materials in physics, Biol. Med. (2020) 1–12.

66 S.C. Karayumak, S. Bouix, L. Ning, A. James, T. Crow, M. Shenton, M. Kubicki, Y. Rathi, Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters, Neuroimage 184 (2019) 180–200.

67 Y. Zhang, G. Parmigiani, W.E. Johnson, ComBat-Seq: batch effect adjustment for RNA-Seq count data, NAR Genom. Bioinform. 2 (2020) lqaa078.

68 C.K. Stein, P. Qu, J. Epstein, A. Buros, A. Rosenthal, J. Crowley, G. Morgan, B. Barlogie, Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat, BMC Bioinformatics 16 (2015) 1–9.

69 R. Da-Ano, I. Masson, F. Lucia, M. Doré, P. Robin, J. Alfieri, C. Rousseau, A. Mervoyer, C. Reinhold, J. Castelli, Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies, Sci. Rep. 10 (2020) 1–12.

70 C. Müller, A. Schillert, C. Röthemeier, D.-.A. Trégouët, C. Proust, H. Binder, N. Pfeiffer, M. Beutel, K.J. Lackner, R.B. Schnabel, Removing batch effects from

71 longitudinal gene expression-quantile normalization plus ComBat as best approach for microarray transcriptome data, PLoS ONE 11 (2016), e0156594.

71 M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C.M. Perou, J.S. Marron, Adjustment of systematic microarray data biases, Bioinformatics 20 (2004) 105–114.

72 A.A. Shabalin, H. Tjelmeland, C. Fan, C.M. Perou, A.B. Nobel, Merging two gene-expression studies via cross-platform normalization, Bioinformatics 24 (2008) 1154–1160.

73 I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-.r. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with Harmony, Nat. Methods 16 (2019) 1289–1296.

74 L. Haghverdi, A.T. Lun, M.D. Morgan, J.C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors, Nat. Biotechnol. 36 (2018) 421–427.

75 B. Hie, B. Bryson, B. Berger, Efficient integration of heterogeneous single-cell transcriptomes using Scanorama, Nat. Biotechnol. 37 (2019) 685–691.

76 F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biol. 19 (2018) 1–5.

77 K. Polański, M.D. Young, Z. Miao, K.B. Meyer, S.A. Teichmann, J.-.E. Park, BBKNN: fast batch alignment of single cell transcriptomes, Bioinformatics 36 (2020) 964–965.

78 L. McInnes, J. Healy, J. Melville, Umap: uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426, (2018).

79 A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat. Biotechnol. 36 (2018) 411–420.

80 J.A. Gagnon-Bartsch, T.P. Speed, Using control genes to correct for unwanted variation in microarray data, Biostatistics 13 (2012) 539–552.

81 D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, J.-.P. Vert, A general and flexible method for signal extraction from single-cell RNA-seq data, Nat. Commun. 9 (2018) 1–17.

82 O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Natl. Acad. Sci. 97 (2000) 10101–10106.

83 Y. Lin, S. Ghazanfar, K.Y. Wang, J.A. Gagnon-Bartsch, K.K. Lo, X. Su, Z.-.G. Han, J. T. Ormerod, T.P. Speed, P. Yang, scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets, Proc. Natl. Acad. Sci. 116 (2019) 9775–9784.

84 J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (2012) 882–883.

85 G.K. Smyth, T. Speed, Normalization of cDNA microarray data, Methods 31 (2003) 265–273.

86 J.-.P. Fortin, E.M. Sweeney, J. Muschelli, C.M. Crainiceanu, R.T. Shinohara, A.s.D. N. Initiative, Removing inter-subject technical variability in magnetic resonance imaging studies, Neuroimage 132 (2016) 198–212.

87 H. Mirzaalian, L. Ning, P. Savadjiev, O. Pasternak, S. Bouix, O. Michailovich, S. Karmacharya, G. Grant, C.E. Marx, R.A. Morey, Multi-site harmonization of diffusion MRI data in a registration framework, Brain Imaging Behav. 12 (2018) 284–295.

88 H. Mirzaalian, A. de Pierrefeu, P. Savadjiev, O. Pasternak, S. Bouix, M. Kubicki, C.-. F. Westin, M.E. Shenton, Y. Rathi, Harmonizing diffusion MRI data across multiple sites and scanners, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 12–19.

89 Y. Zhang, D.F. Jenkins, S. Manimaran, W.E. Johnson, Alternative empirical Bayes models for adjusting for batch effects in genomic studies, BMC Bioinformatics 19 (2018) 1–15.

90 K.M. Huynh, G. Chen, Y. Wu, D. Shen, P.-.T. Yap, Multi-site harmonization of diffusion MRI data via method of moments, IEEE Trans. Med. Imaging 38 (2019) 1599–1609.

91 J. Wrobel, M. Martin, R. Bakshi, P. Calabresi, M. Elliot, D. Roalf, R. Gur, R. Gur, R. Henry, G. Nair, Intensity warping for multisite MRI harmonization, Neuroimage 223 (2020), 117242.

92 A. Llera, I. Huertas, P. Mir, C.F. Beckmann, Quantitative intensity harmonization of dopamine transporter SPECT images using gamma mixture models, Mol. Imaging Biol. 21 (2019) 339–347.

93 C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, D.Y.W. Solís, C. Molter, R. Duque, H. Bersini, A. Nowé, GENESHIFT: a nonparametric approach for integrating microarray gene expression data based on the inner product as a distance measure between the distributions of genes, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2013) 383–392.

94 D. Mackin, X. Fave, L. Zhang, J. Yang, A.K. Jones, C.S. Ng, L. Court, Harmonizing the pixel size in retrospective computed tomography radiomics studies, PLoS ONE 12 (2017), e0178524.

95 H. Moradmand, S.M.R. Aghamiri, R. Ghaderi, Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma, J. Appl. Clin. Med. Phy. 21 (2020) 179–190.

96 I. Pitas, Digital Image Processing Algorithms and Applications, John Wiley & Sons, 2000.

97 R.T. Shinohara, E.M. Sweeney, J. Goldsmith, N. Shiee, F.J. Mateen, P.A. Calabresi, S. Jarso, D.L. Pham, D.S. Reich, C.M. Crainiceanu, Statistical normalization techniques for magnetic resonance imaging, NeuroImage Clin. 6 (2014) 9–19.

98 E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graph. Appl. 21 (2001) 34–41.

99 C.P. Loizou, V. Murray, M.S. Pattichis, I. Seimenis, M. Pantziaris, C.S. Pattichis, Multiscale amplitude-modulation frequency-modulation (AM–FM) texture analysis

of multiple sclerosis in brain MRI images, IEEE Trans. Inf. Technol. Biomed. 15 (2010) 119–129.

[100] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D.L. Arnold, D.L. Collins, T. Arbel, Evaluating intensity normalization on MRIs of human brain with multiple sclerosis, Med. Image Anal. 15 (2011) 267–282.

[101] S. Roy, S. Lal, J.R. Kini, Novel color normalization method for Hematoxylin & Eosin stained histopathology images, IEEE Access 7 (2019) 28982–28998.

[102] M.D. Zarella, C. Yeoh, D.E. Breen, F.U. Garcia, An alternative reference space for H&E color normalization, PLoS ONE 12 (2017), e0174489.

[103] X. Li, K.N. Plataniotis, A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics, IEEE Trans. Biomed. Eng. 62 (2015) 1862–1873.

[104] D.F. Swinehart, The beer-lambert law, J. Chem. Educ. 39 (1962) 333.

[105] T.A.A. Tosta, P.R. de Faria, L.A. Neves, M.Z. do Nascimento, Color normalization of faded H&E-stained histological images using spectral matching, Comput. Biol. Med. 111 (2019), 103344.

[106] A.M. Khan, N. Rajpoot, D. Treanor, D. Magee, A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution, IEEE Trans. Biomed. Eng. 61 (2014) 1729–1738.

[107] A.C. Ruifrok, D.A. Johnston, Quantification of histochemical staining by color deconvolution, Anal. Quant. Cytol. Histol. 23 (2001) 291–299.

[108] M.Z. Hoque, A. Keskinarkaus, P. Nyberg, T. Seppänen, Retinex model based stain normalization technique for whole slide image analysis, Comput. Med. Imaging Graph. 90 (2021), 101901.

[109] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, N. Navab, Structure-preserving color normalization and sparse stain separation for histological images, IEEE Trans. Med. Imaging 35 (2016) 1962–1971.

[110] G. Lei, Y. Xia, D.-.H. Zhai, W. Zhang, D. Chen, D. Wang, StainCNNs: an efficient stain feature learning method, Neurocomputing 406 (2020) 267–273.

[111] Y. Zheng, Z. Jiang, H. Zhang, F. Xie, J. Shi, C. Xue, Adaptive color deconvolution for histological WSI normalization, Comput. Methods Programs Biomed. 170 (2019) 107–120.

[112] P. Maji, S. Mahapatra, Rough-fuzzy circular clustering for color normalization of histological images, Fundam. Inform. 164 (2019) 103–117.

[113] P. Maji, S. Mahapatra, Circular clustering in fuzzy approximation spaces for color normalization of histological images, IEEE Trans Med Imaging 39 (2019) 1735–1745.

[114] H.-.D. Cheng, X. Cai, R. Min, A novel approach to color normalization using neural network, Neural Comput. Appl. 18 (2009) 237–247.

[115] V. Golkov, A. Dosovitskiy, J.I. Sperl, M.I. Menzel, M. Czisch, P. Sämann, T. Brox, D. Cremers, Q-space deep learning: twelve-fold shorter and model-free diffusion MRI scans, IEEE Trans. Med. Imaging 35 (2016) 1344–1351.

[116] S. Koppers, L. Bloy, J.I. Berman, C.M. Tax, J.C. Edgar, D. Merhof, Spherical harmonic residual network for diffusion signal harmonization, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 173–182.

[117] S.C. Karayumak, M. Kubicki, Y. Rathi, Harmonizing diffusion MRI data across magnetic field strengths, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 116–124.

[118] B.E. Dewey, C. Zhao, J.C. Reinhold, A. Carass, K.C. Fitzgerald, E.S. Sotirchos, S. Saidha, J. Oh, D.L. Pham, P.A. Calabresi, DeepHarmony: a deep learning approach to contrast harmonization across scanner changes, Magn. Reson. Imaging 64 (2019) 160–170.

[119] Q. Tong, T. Gong, H. He, Z. Wang, W. Yu, J. Zhang, L. Zhai, H. Cui, X. Meng, C. W. Tax, A deep learning–based method for improving reliability of multicenter diffusion kurtosis imaging with varied acquisition protocols, Magn. Reson. Imaging 73 (2020) 31–44.

[120] S. Park, S.M. Lee, K.-.H. Do, J.-.G. Lee, W. Bae, H. Park, K.-.H. Jung, J.B. Seo, Deep learning algorithm for reducing CT slice thickness: effect on reproducibility of radiomic features in lung cancer, Korean J. Radiol. 20 (2019) 1431–1440.

[121] U. Shaham, K.P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, Y. Kluger, Removal of batch effects using distribution-matching residual networks, Bioinformatics 33 (2017) 2539–2546.

[122] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[123] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International conference on machine learning, PMLR, 2015, pp. 448–456.

[124] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, A kernel method for the two-sample-problem, Adv. Neural Inf. Process. Syst. 19 (2006) 513–520.

[125] A. Jog, A. Carass, S. Roy, D.L. Pham, J.L. Prince, MR image synthesis by contrast learning on neighborhood ensembles, Med. Image Anal. 24 (2015) 63–76.

[126] A. Jog, A. Carass, S. Roy, D.L. Pham, J.L. Prince, Random forest regression for magnetic resonance image synthesis, Med. Image Anal. 35 (2017) 475–488.

[127] J.-.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[128] F. Zhao, Z. Wu, L. Wang, W. Lin, S. Xia, D. Shen, G. Li, Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 475–483.

[129] M. Ren, N. Dey, J. Fishbaugh, G. Gerig, Segmentation-renormalized deep feature modulation for unpaired image harmonization, IEEE Trans. Med. Imaging 40 (2021) 1519–1530.

[130] J. Zhong, Y. Wang, J. Li, X. Xue, S. Liu, M. Wang, X. Gao, Q. Wang, J. Yang, X. Li, Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development, Biomed. Eng. Online 19 (2020) 1–18.

[131] D. Moyer, G. Ver Steeg, C.M. Tax, P.M. Thompson, Scanner invariant representations for diffusion MRI harmonization, Magn. Reson. Med. 84 (2020) 2174–2189.

[132] N. Russkikh, D. Antonets, D. Shtokalo, A. Makarov, Y. Vyatkin, A. Zakharov, E. Terentyev, Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis, Bioinformatics 36 (2020) 5076–5085.

[133] N. Johansen, G. Quon, scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data, Genome Biol. 20 (2019) 1–21.

[134] P. Haeusser, A. Mordvintsev, D. Cremers, Learning by association–a versatile semi-supervised training method for neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 89–98.

[135] D. Wang, S. Hou, L. Zhang, X. Wang, B. Liu, Z. Zhang, iMAP: integration of multiple single-cell datasets by adversarial paired transfer networks, Genome Biol. 22 (2021) 1–24.

[136] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding, J. Machine Learn. Res. 11 (2010).

[137] S. St-Jean, P. Coupé, M. Descoteaux, Non local spatial and angular matching: enabling higher spatial resolution diffusion MRI datasets through adaptive denoising, Med. Image Anal. 32 (2016) 115–130.

[138] T.A.A. Tosta, P.R. de Faria, J.P.S. Servato, L.A. Neves, G.F. Roberto, A.S. Martins, M. Z. do Nascimento, Unsupervised method for normalization of hematoxylin-eosin stain in histological images, Comput. Med. Imaging Graph. 77 (2019), 101646.

[139] C. Lu, J. Shi, J. Jia, Online robust dictionary learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 415–422.

[140] X. Li, K. Wang, Y. Lyu, H. Pan, J. Zhang, D. Stambolian, K. Susztak, M.P. Reilly, G. Hu, M. Li, Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis, Nat. Commun. 11 (2020) 1–14.

[141] V.D. Blondel, J.-.L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. (2008) P10008, 2008.

[142] T. Wang, T.S. Johnson, W. Shao, Z. Lu, B.R. Helm, J. Zhang, K. Huang, BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes, Genome Biol. 20 (2019) 1–15.

[143] H. Guan, Y. Liu, E. Yang, P.-.T. Yap, D. Shen, M. Liu, Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification, Med. Image Anal. 71 (2021), 102076.

[144] N.K. Dinsdale, M. Jenkinson, A.I. Namburete, Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal, Neuroimage 228 (2021), 117689.

[145] S. Ge, H. Wang, A. Alavi, E. Xing, Z. Bar-Joseph, Supervised adversarial alignment of single-cell RNA-seq data, J. Comput. Biol. 28 (2021) 501–513.

[146] Z. Rong, Q. Tan, L. Cao, L. Zhang, K. Deng, Y. Huang, Z.-.J. Zhu, Z. Li, K. Li, NormAE: deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data, Anal. Chem. 92 (2020) 5082–5090.

[147] M. Büttner, Z. Miao, F.A. Wolf, S.A. Teichmann, F.J. Theis, A test metric for assessing single-cell RNA-seq batch correction, Nat. Methods 16 (2019) 43–49.

[148] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004).

[149] A. Kolaman, O. Yadid-Pecht, Quaternion structural similarity: a new quality index for color images, IEEE Trans. Image Process. 21 (2011) 1526–1536.

[150] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: a feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (2011) 2378–2386.

[151] J.-.F. Pambrun, R. Noumeir, Limitations of the SSIM quality metric in the context of diagnostic imaging, in: Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 2960–2963.

[152] L.G. Nyúl, J.K. Udupa, X. Zhang, New variants of a method of MRI scale standardization, IEEE Trans. Med. Imaging 19 (2000) 143–150.

[153] A. Albert, L. Zhang, A novel definition of the multivariate coefficient of variation, Biomet. J. 52 (2010) 667–675.

[154] P. Chirra, P. Leo, M. Yim, B.N. Bloch, A.R. Rastinehad, A. Purysko, M. Rosen, A. Madabhushi, S.E. Viswanath, Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI, J. Med. Imaging 6 (2019), 024502.

[155] I. Lawrence, K. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics (1989) 255–268.

[156] D. Liljequist, B. Elfving, K. Skavberg Roaldsen, Intraclass correlation–a discussion and demonstration of basic features, PLoS ONE 14 (2019), e0219854.

[157] F. Orlhac, A. Lecler, J. Savatovski, J. Goya-Outi, C. Nioche, F. Charbonneau, N. Ayache, F. Frouin, L. Duron, I. Buvat, How can we combat multicenter variability in MR radiomics? Validation of a correction procedure, Eur. Radiol. 31 (2021) 2272–2280.

[158] R. Mahon, M. Ghita, G. Hugo, E. Weiss, ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets, Phy. Med. Biol. 65 (2020), 015010.

[159] G.S. Ioannidis, E. Trivizakis, I. Metzakis, S. Papagiannakis, E. Lagoudaki, K. Marias, Pathomics and deep learning classification of a heterogeneous fluorescence histology image dataset, Appl. Sci. 11 (2021) 3796.

160 M. Lotfollahi, F.A. Wolf, F.J. Theis, scGen predicts single-cell perturbation responses, Nat. Methods 16 (2019) 715–721.

161 X. Jiang, G.-.B. Bian, Z. Tian, Removal of artifacts from EEG signals: a review, Sensors 19 (2019) 987.

162 P. He, M. Kahle, G. Wilson, C. Russell, Removal of ocular artifacts from EEG: a comparison of adaptive filtering method and regression method using simulated data, in: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, IEEE, 2006, pp. 1110–1113.

163 P.S. Kumar, R. Arumuganathan, K. Sivakumar, C. Vimal, Removal of ocular artifacts in the EEG through wavelet transform without using an EOG reference channel, Int. J. Open Problems Compt. Math 1 (2008) 188–200.

164 H.J. Aerts, E.R. Velazquez, R.T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (2014) 1–9.

165 A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inform. Fusion 58 (2020) 82–115.

166 G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond, Inform. Fusion 77 (2022) 29–52.

167 A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inform. Fusion 79 (2022) 263–278.

168 F.J. López-González, J. Silva-Rodríguez, J. Paredes-Pacheco, A. Niñerola-Baizán, N. Efthimiou, C. Martín-Martín, A. Moscoso, Á. Ruibal, N. Roé-Vellvé, P. Aguiar, Intensity normalization methods in brain FDG-PET quantification, Neuroimage 222 (2020), 117229.

[169] Mongan, John, Linda Moy, and Charles E. Kahn Jr. "Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers." *Radiology: Artificial Intelligence* 2.2 (2020): e200029.

[170] S. St-Jean, M.A. Viergever, A. Leemans, Harmonization of diffusion MRI data sets with adaptive dictionary learning, Human brain mapping 41 (16) (2020) 4478–4499.