BEEx Is an Open-Source Tool That Evaluates Batch Effects in Medical Images to Enable Multicenter Studies

Yuxin Wu[1,2,3], Xiongjun Xu[4], Yuan Cheng[1,2,3], Xiuming Zhang[5], Fanxi Liu[6], Zhenhui Li[7], Lei Hu[1,2,3], Anant Madabhushi[8,9,10], Peng Gao[11,12], Zaiyi Liu[1,3], and Cheng Lu[1,2,3]

ABSTRACT

The batch effect is a nonbiological variation that arises from technical differences across different batches of data during the data generation process for acquisition-related reasons, such as collection of images at different sites or using different scanners. This phenomenon can affect the robustness and generalizability of computational pathology- or radiology-based cancer diagnostic models, especially in multicenter studies. To address this issue, we developed an open-source platform, Batch Effect Explorer (BEEx), that is designed to qualitatively and quantitatively determine whether batch effects exist among medical image datasets from different sites. A suite of tools was incorporated into BEEx that provide visualization and quantitative metrics based on intensity, gradient, and texture features to allow users to determine whether there are any image variables or combinations of variables that can distinguish datasets from different sites in an unsupervised manner. BEEx was designed to support various medical imaging techniques, including microscopy and radiology. Four use cases clearly demonstrated the ability of BEEx to identify batch effects and validated the effectiveness of rectification methods for batch effect reduction. Overall, BEEx is a scalable and versatile framework designed to read, process, and analyze a wide range of medical images to facilitate the identification and mitigation of batch effects, which can enhance the reliability and validity of image-based studies.

Significance: BEEx is a prescreening tool for image-based analyses that allows researchers to evaluate batch effects in multicenter studies and determine their origin and magnitude to facilitate development of accurate AI–based cancer models.

Introduction

Batch effects are nonbiological variations arising from technical differences across different batches of data during the data generation process, such as in images curated from different sites or obtained using different scanners. These acquisition-related factors impact the imaging data in a manner that impedes the ability of machine learning algorithms to generalize to data from new sites. In the literature, researchers showed that they can use machine learning classifiers to predict the originating sites of data (1), which indicates that there is abundant site-related batch effect information embedded in the image data. The causes of batch effects are varied. In digital pathology, batch effects can arise from the use of different scanners, the method of image preprocessing, differences in operator handling, and the manner in which the slides are prepared. In radiology, batch effects can arise from scanner and protocol variability, acquisition parameters, and image preprocessing methods.

Batch effects pose substantial obstacles in the field of medical image analysis, particularly with regard to machine learning applications. When algorithms are trained on one batch or cohort, they may not perform well when applied to another and may lack robustness or generalizability. Previously, researchers have demonstrated the extent and impact of batch effects in various contexts. Kothari and colleagues (2) demonstrated that batch effects can change morphologic image features and decrease the prediction performance in multi-batch studies. Stopsack and colleagues (3) demonstrated that in half of the 20 protein biomarkers obtained from 1,448 men in two nationwide cohort studies, more than 10% of the biomarker variance was attributable to batch differences. Their study proved that the differences between different batches could be

[1]Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China. [2]Medical Research Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China. [3]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou, China. [4]Department of Stomatology, The Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China. [5]Department of Pathology, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China. [6]Department of Computer Science, School of Computing, National University of Singapore, Singapore, Republic of Singapore. [7]Department of Radiology, The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Yunnan Cancer Center, Kunming, China. [8]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia. [9]Radiology and Imaging Sciences, Biomedical Informatics (BMI) and Pathology, Georgia Institute of Technology and Emory University, Atlanta, Georgia. [10]Atlanta Veterans Administration Medical Center, Atlanta, Georgia. [11]Key Laboratory for Experimental Teratology of Ministry of Education, Department of Pathology, School of Basic Medical Sciences, Shandong University, Jinan, China. [12]Department of Pathology, Qilu Hospital, Shandong University, Jinan, China.

Y. Wu, X. Xu, Y. Cheng, and X. Zhang are co-first authors and contributed equally to this article.

Corresponding Authors: Cheng Lu, Department of Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, 106 Zhongshaner Road, Yuexiu District, Guangzhou 510080, Guangdong Province, China. E-mail: lucheng@gdph.org.cn; Zaiyi Liu, E-mail: liuzaiyi@gdph.org.cn; and Peng Gao, Department of Pathology, Qilu Hospital, Shandong University, 107 Wenhua Xilu, Jinan 250012, Shandong Province, China. E-mail: gaopeng@sdu.edu.cn

Cancer Res 2025;85:218–30

doi: 10.1158/0008-5472.CAN-23-3846

attributed more to batch effects than to different patient or tumor characteristics. An increasing number of researchers have focused on multicenter studies (4, 5) to validate the robustness and generalizability of artificial intelligence (AI) models. Therefore, there is an urgent need for efficient tools that can assess batch effects in multicenter settings.

In the context of genomic analysis, Zhu and colleagues (6) presented BatchServer, an open-source R/Shiny-based platform for batch effect analysis on proteomic and transcriptomic data. However, the platform is incompatible with medical imaging. Stopsack and colleagues (3) demonstrated the influence of batch effects on tissue microarrays (TMA) and used diverse algorithms to alleviate them. A quantitative quality control tool, MRQy (7), was specifically designed for MRI data quality control and can quantify site- or scanner-specific variations in image resolution or contrast as well as imaging artifacts such as noise or inhomogeneity. Similarly, for histologic images, HistoQC (8), a tool for rapid quality control, has been proposed to delineate artifacts and identify cohort-level outliers. Chen and colleagues (9) utilized HistoQC-generated image variables to train a random forest to classify the originating sites of whole-slide images (WSI), and found batch effects in at least three laboratories. However, to the best of our knowledge, no open-source, preanalytic platform has been specifically designed to explore image-based batch effects on medical images obtained from multiple centers. Therefore, to enable researchers to perform batch effect analysis prior to the validation of AI models in a multicenter setting, we present the open-source Batch Effect Explorer (BEEx), which is implemented via Python and is compatible with a wide range of types of medical images.

Materials and Methods

Image data for multicenter studies

To prove that modern AI diagnostic and prognostic models are robust and generalizable, many studies are currently testing these models on multicenter datasets. In addition, researchers are not only focusing on a single modality but also combining information from different modalities. To demonstrate the effectiveness of BEEx for analyzing the batch effect among different datasets and modalities, we prepared four multicenter case studies, as shown in **Fig. 1A**, each of which focuses on a specific type of medical image: WSI, TMA, MRI, and CT. These image types are widely used in pathologic and radiologic studies and play a crucial role in clinical diagnostics, treatment planning, and biomedical research.

Case study: WSI

This case study examined four distinct hepatocellular carcinoma (HCC) WSI cohorts from different medical institutions. These cohorts included 162 images of 156 patients from Guangdong Provincial People's Hospital (GDPH), 290 images of 274 patients from The Cancer Genome Atlas (TCGA), 190 images of 182 patients from The First Affiliated Hospital of Kunming Medical University (FAHKU), and 167 images of 166 patients from The First Affiliated Hospital Zhejiang University School of Medicine (FAHZU).

Case study: virtual TMA

This case study involved three virtual TMA (VTMA) cohorts consisting of 293 patients with breast cancer and 293 images: VTMA-1 (102 patients and 102 images), VTMA-2 (94 patients and 94 images), and VTMA-3 (97 patients and 97 images). These VTMA spots were digitally extracted from the tumor regions of three resected tissue WSI cohorts: Shandong University Qilu Hospital, Qingdao University Hospital, and the Second Hospital of Shandong University. Note that all histologic sections were removed from preserved tissues from different medical institutes. The recut tissues were processed, stained in different laboratories while using the same reagents and antibodies, and subsequently scanned using the same scanner.

Case study: MRI-A and -B

This case study focused on prostate cancer MRIs and drew data from three distinct institutes: Shanghai Sixth People's Hospital (SSPH), which contributed 100 images from 100 patients; Renming Hospital of Wuhan University (RHWU), which provided 100 images from 100 patients; and Yichang Central People's Hospital (YCPH), which supplied 77 images from 77 patients. To determine whether BEEx can detect the reduction in batch effects among datasets processed using batch effect rectification methods and evaluate the downstream task performance after batch effect rectification (10), we investigated case study MRI-B based on case study MRI-A with batch effect rectification applied.

Case study: CT

This was a case–control study with three pseudo cohorts: GDPH1, GDPH2, and GDPH3. These pseudo cohorts were artificially created by randomly selecting cases from the same abdominal CT urography cohort from GDPH. In this case study, each pseudo cohort comprised 100 images from 100 patients.

Ethics Statement

This retrospective study was conducted in accordance with the ethical guidelines of the Declaration of Helsinki. Ethical approvals were obtained from the following institutional review boards: GDPH (KY2023-357-02), FAHKU (KYLX2024-171), and FAHZU (KY2024-0376) for the WSI datasets; the School of Basic Medical Sciences of Shandong University (ECSBMSSDU2023-1-52) for the VTMA datasets; SSPH [2022-KY-073(K)] for the MRI datasets; and GDPH (KY2023-146-01) for the CT dataset. All institutional review boards waived the requirement for informed consent from patients due to the retrospective nature of the study.

Blinding

All the patients remain anonymous in this study.

BEEx

The BEEx platform includes three interconnected modules: (i) Preprocessor: This module is responsible for loading and preprocessing images, segmentation masks, and clinical data and preparing them for subsequent analyses. (ii) Feature Extractor: This module focuses on extracting image intensity-, gradient-, and texture-based features from images for downstream analyses. (iii) Analyzer: This module provides batch effect assessment and visualization. It also includes a suite of tools for various analyses, as shown in **Fig. 1B**, including image overviewer, distribution analysis with violin plots (11), Uniform Manifold Approximation and Projection (UMAP; RRID: SCR_018217; ref. 12), hierarchical clustering (RRID: SCR_014673; ref. 13), principal variation component analysis (PVCA; RRID: SCR_001356; ref. 14), and the proposed batch effect score (BES).

Preprocessor

The preprocessor helps prepare and process data for the feature extractor and analyzer. BEEx supports a wide range of image types and

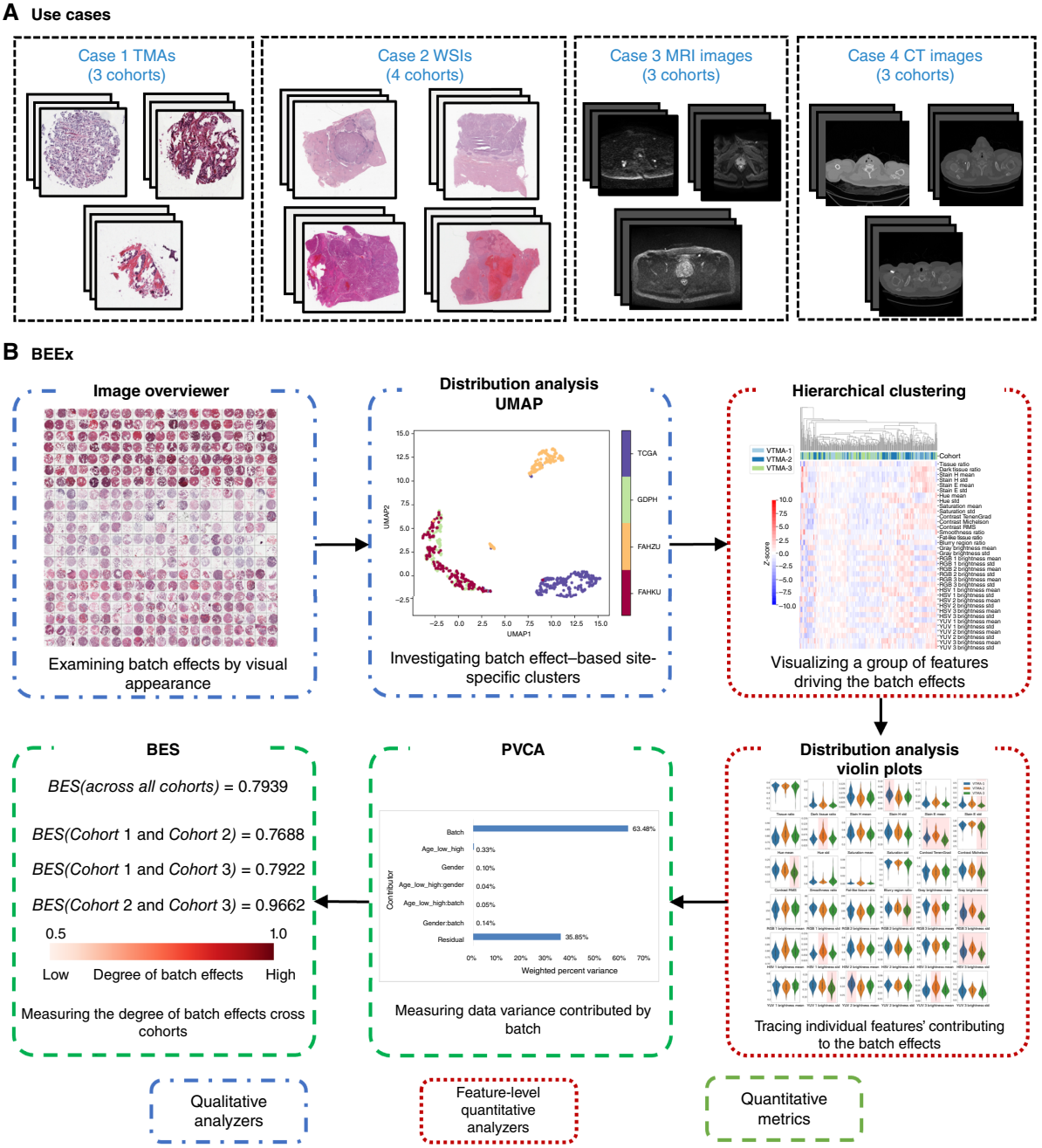

Figure 1.

Overview of the BEEx platform. **A,** Four case studies that include TMA, WSI, MRI, and CT images. **B,** Illustration of the interpretative flowchart of the BEEx platform, including different analyzers with different purposes: (i) image overviewer directly examines batch effects by visual appearance, (ii) distribution analysis with UMAP investigates batch effect–based site-specific clusters, (iii) hierarchical clustering visualizes group of features driving the clusters using a heatmap, (iv) distribution analysis with violin plots helps reveal and trace features contributing to the batch effect, (v) PVCA numerically partitions the variance in data represented by batch and other clinical-related factors, and (vi) BES measures the degree of batch effects across cohorts.

formats for medical images, e.g., NIfTI (.nii.gz), DICOM (.dcm), portable network graphics (.png), and OpenSlide (.svs). In addition, BEEx provides helper scripts to assist users in converting DICOM files into the NIfTI format. Presegmented region of interest (ROI) masks and clinical data are optional during preprocessing. If ROI masks are specified, BEEx analyzes batch effects only within the areas of interest indicated by the ROI masks. Clinical data, if available, are used to perform PVCA.

Feature extractor

In this module, BEEx extracts various intensity-, gradient-, and texture-based features from the processed images. The extraction of these image features was inspired by two quality check toolkits MRQy (RRID: SCR_025779; ref. 7) and HistoQC (RRID: SCR_025780; ref. 8). For digital pathology images, BEEx extracts $n = 36$ features, including color histograms, brightness, and contrast. For radiology images, BEEx extracts at most $n = 27$ features, including the signal-to-noise ratio, mean of the foreground intensity, and entropy focus criterion. A detailed description of each feature is provided in Supplementary Tables S1 and S2.

BEEx analyzers

BEEx uses six analyzers to visualize and evaluate batch effects from different perspectives: image overviewer, distribution analysis with UMAP, hierarchical clustering, distribution analysis with violin plots, PVCA, and the proposed BES. Specifically, image overviewer and distribution analysis with UMAP evaluate the existence of batch effects through direct visual examination. Hierarchical clustering and distribution analysis with violin plots explore batch effects on feature level. PVCA and BES quantitatively measure the degree of batch effect.

Image overviewer directly examines batch effects based on visual appearance. It resizes raw images and arranges them into one single image plot such that the visual differences and characteristics of the image (e.g., color, style, and shape) from different cohorts become more distinct and easier to examine. A flowchart of image overviewer is shown in Supplementary Fig. S1.

Distribution analysis with UMAP projects image features from different cohorts into a two-dimensional space for better batch effect examination (12). By reducing the dimensionality of the data, we can visualize complex patterns and structures more easily and check for site-specific clusters. If a batch effect is present, distinguishable clusters are present in the plot. The pseudo code of distribution analysis with UMAP is provided in Supplementary Algorithm S1.

Hierarchical clustering, also known as clustergram, further interprets the clustering results of distribution analysis with UMAP. It creates a hierarchy of clusters and visualizes the features driving the clusters using a heatmap (13). The resulting clusters are visualized using a heatmap with color gradients to represent the numerical values of the image features, which allows us to visualize the relationships, similarities, and differences between different image samples. If a batch effect exists, the samples in the analyzer plot are primarily grouped based on their cohort. The pseudo code for hierarchical clustering is given in Supplementary Algorithm S2.

Distribution analysis with violin plots illustrates the distribution of image features calculated by the feature extractor for different cohorts, which helps to reveal and trace the potential batch effect of different cohorts (11). That is, to determine which or what kinds of features contribute to batch effects, this analyzer performs the Wilcoxon rank-sum test in a pairwise manner (15) on the extracted image features from different cohorts. The Wilcoxon rank-sum test is a nonparametric statistical hypothesis test used to compare independent samples (i.e., image features from different cohorts in our case) and assess whether their distributions differ significantly. If the feature values of one cohort are significantly different from those of the other cohorts ($P < 0.05$), this cohort is highlighted in red in the sub-violin plots. Thus, users can assess the degree of the batch effect not only at the overall level but also at the cohort level. Moreover, tracing the features related to batch effects may help users better understand the extent of variability in their data and enlighten them on how batch effects can be alleviated.

PVCA combines PCA and VCA to numerically partition the variance in data represented by batch and other clinical-related factors into different major components (14). This can be useful for identifying and quantifying major sources of variability in the data. If batch effects exist, the batch variance in the plot is one of the main contributors to the overall variation. The implementation of PVCA is shown in Supplementary Algorithm S3.

BES serves as another numeric index of the degree of batch effects. It is a measure of the batch effect through the assessment of the stability of sample clustering in cohorts based on consensus clustering (16). We assumed that cohorts with a high degree of batch effects contained samples that were stably clustered in multiple clustering runs. Thus, given resampling iterations H, the sampled datasets can be defined as $D = \{D^1, D^2, \cdots, D^H\}$ with a predefined proportion, e.g., 80%. In each resampling iteration $h \in \{1, 2, \cdots, H\}$, let $M^{(h)}$ denote the $(N \times N)$ connectivity matrix corresponding to dataset $D^{(h)}$, in which N is the number of all samples. The results of clustering of pairs of samples can be defined as entries of the following matrix:

$$M^{(h)}(i,j) = \begin{cases} 1 \textit{ if item i and j belong to the same cluster} \\ 0 \textit{ otherwise} \end{cases} \quad \text{(A)}$$

Let $I^{(h)}$ be the $(N \times N)$ indicator matrix. Here, $I^{(h)}(i,j)$ equals 1 when items i and j are both present in the sampled dataset $\boldsymbol{D^{(h)}}$. Otherwise, it equals 0. The consensus matrix $\mathscr{M}$ can be defined as the normalized sum of the connectivity matrices of all the resampled datasets:

$$\mathscr{M}(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)} \quad \text{(B)}$$

In other words, the consensus matrix $\mathscr{M}(i,j)$ records the number of times items i and j are assigned to the same cluster, divided by the total number of times both items are selected in the same resampled datasets. Therefore, the consensus matrix measures how stably and robustly the sample pairs are clustered together.

Let $C(i,j)$ indicate whether items i and j belong to the same cohort; we then define it as follows:

$$C(i,j) = \begin{cases} 1 \textit{ if item i and j belong to the same cohort} \\ 0 \textit{ otherwise} \end{cases} \quad \text{(C)}$$

Here, we define the BES as the AUC (17) of the consensus matrix $\mathscr{M}$ and the cohort matrix C, in which item $i < j$:

$$S = \{(i,j) \mid 1 \le i < j \le N\} \quad \text{(D)}$$

$$\mathscr{M}' = \{\mathscr{M}(i,j) \mid (i,j) \in S\} \quad \text{(E)}$$

$$C' = \{C(i,j) \mid (i,j) \in S\} \quad \text{(F)}$$

$$BES = AUC\left(C', \mathscr{M}'\right) \quad \text{(G)}$$

When the value of the BES approaches 1, the samples are easily and stably clustered according to their cohorts, indicating a high degree of batch effects among their cohorts. In contrast, if the value of the BES is close to 0.5, the samples are randomly clustered, indicating a low degree of batch effects among the cohorts.

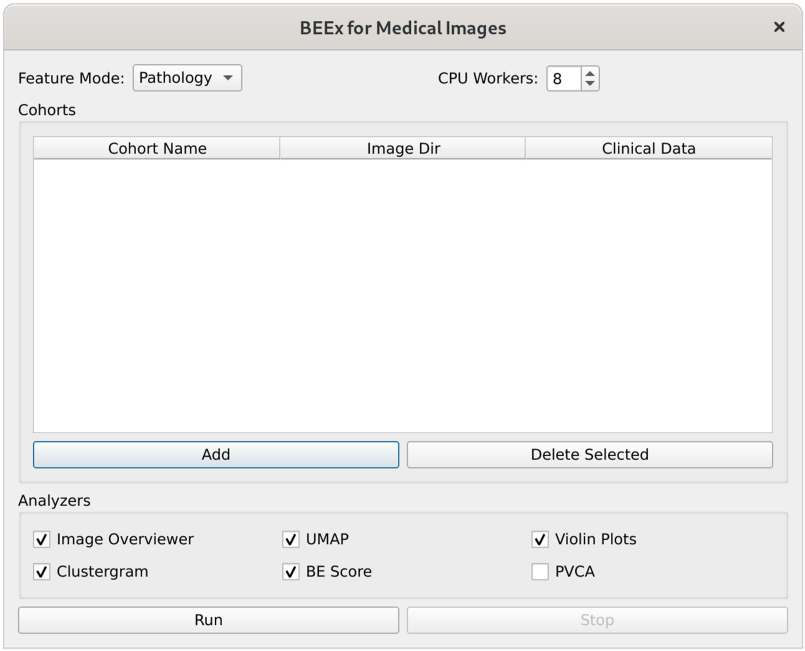

Figure 2.
Screenshot of the BEEx GUI. Users need to specify the location of the image files of each cohort with a few clicks. After clicking the "Run" button with the selected analyzers, the results pop up in separate windows.

If the BES value is below 0.5, it may indicate data issues within the dataset, such as mistakenly treating data from a single center as if they were from different centers or significant disparities in the number of samples from different centers. In such cases, users should review their data and consider balancing the sample sizes to address these issues. Note that the BES can be used not only for all cohorts in a dataset but also for cohort levels within a multicenter dataset.

Getting started with BEEx

Easily runnable graphical user interface (GUI) versions of BEEx for Linux and Windows can be found at https://github.com/wuusn/beex, along with the sample data used in this study for quick reproduction. **Figure 2** shows the BEEx GUI.

Technical requirements: a 64-bit operating system and at least 16 GB of RAM. However, a larger RAM size is recommended for WSI processing and for large numbers of images. Supplementary Figure S2 illustrates how the time cost of BEEx varies with different numbers of centers and images.

BEEx supports a large number of medical image formats, including basic formats (.png, .jpg, and .tiff), OpenSlide (.svs), NIfTI (.nii.gz), and DICOM (.dcm).

All BEEx codes are open-source and can be found in our GitHub repository (https://github.com/wuusn/beex) along with current issues and proposed code changes. **Figure 3** shows the use of BEEx from the command line. Supplementary Figures S3 to S8 provide a step-by-step tutorial on how to use the BEEx GUI.

If you need help, feel free to open an issue in the GitHub repository.

Software environment

BEEx requires a 64-bit Windows, Linux, or Mac operating system with at least 16 GB of RAM. It is implemented using a Python 3.11.5 programming environment with the openslide-python 1.30, pillow 10.0.1, MedPy 0.4.0, and pydicom 2.4.3 (RRID: SCR_002573) for medical image reading; NumPy 1.25.2 (RRID: SCR_008633), pandas 2.1.1, scikit-image 0.21.0, and scikit-learn 1.3.1 for data processing and analysis; and matplotlib 3.8.0 and seaborn 0.13.0 for visualization. The BEEx GUI was based on PyQt and PySide6 6.7.1.

Data availability

The WSI, VTMA, MRI, and CT data analyzed in this study are publicly available at https://doi.org/10.6084/m9.figshare.25271215.v1. The TCGA cohort of WSI data analyzed in this study was obtained from TCGA at https://portal.gdc.cancer.gov/projects/TCGA-LIHC.

The Python code for the BEEx framework is available on GitHub at https://github.com/wuusn/beex. A reproducible capsule (https://codeocean.com/capsule/2528710/tree/v1) is hosted on the Code Ocean platform.

Results

In this study, we demonstrated the use of BEEx through four case studies, each of which focused on a specific type of medical image. In case study WSI, the average image sizes at 40× magnification for GDPH, TCGA, FAHKU, and FAHZU were 77,492 × 79,768, 96,965 × 68,054, 100,825 × 80,537, and 84,246 × 64,355 pixels, respectively. The average ages within these cohorts were 55 ± 11, 60 ± 12, 52 ± 11, and 55 ± 10, respectively. The male-to-female ratios in these cohorts were (85.3% and 14.7%), (68.6% and 31.4%), (91.2% and 8.8%), and (81.9% and 18.1%), respectively. In case study VTMA, each TMA image had a size of 2,800 × 2,800 pixels at ×40 magnification. VTMA-1, VTMA-2, and VTMA-3 contained 19 noninvasive and 83 invasive tumor images, 1 noninvasive and 93 invasive tumor images, and 45 noninvasive and 52 invasive tumor images, respectively. Additionally, they contained 7 low–, 51 medium–, and 44 high–histologic-grade tumor images; 8 low–, 48 medium–, and 38 high–histologic-grade tumor images; and 13 low–, 42 medium–, and 42 high–histologic-grade tumor images, respectively. In case study MRI, the average ages within SSPH,

Example configuration file example.yaml

```
CaseTMA:
    feature_mode: path
    n_workers: 8
    cohort_dir:
        - /path/to/SUQH/imgs
        - /path/to/QDUH/imgs
        - /path/to/SHSU/imgs
    cohort_name:
        - VTMA-1
        - VTMA-2
        - VTMA-3
    clinical_data:
        - /mnt/hd0/project_large_files/bee/VTMA_v3/SUQH/clinical.xlsx
        - /mnt/hd0/project_large_files/bee/VTMA_v3/QDUH/clinical.xlsx
        - /mnt/hd0/project_large_files/bee/VTMA_v3/SHSU/clinical.xlsx
    clinical_column:
        - Invasion
        - Overgrade
    save_dir: /data/project_large_files/bee/deploy_test/CaseVTMA
```

Python run command

```
$ python bee.py example.yaml
```

Figure 3.

Example of running the code from the command line. The user needs to create a configuration .yaml file indicating the basic information about the dataset and then run the Python script with the .yaml file as the input.

RHWU, and YCPH were 71 ± 9, 71 ± 8, and 69 ± 9, respectively. Additionally, the average values of the PSA test were 20.65 ± 20.77, 75.64 ± 146.14, and 15.40 ± 24.57, respectively. The imaging matrices for these cohorts were 178 × 132, 256 × 256, and 224 × 224 pixels, respectively. In case study CT, the average ages of GDPH1, GDPH2, and GDPH3 were 62 ± 12, 61 ± 11, and 64 ± 12, respectively. The male-to-female ratios in these cohorts were (68.37% and 31.63%), (58.16% and 41.84%), and (60% and 40%), respectively. All the imaging matrices for these cohorts were 512 × 512. In case study VTMA, the results of BEEx demonstrated that there was a low degree of batch effects. Conversely, in case study WSI and case study MRI-A and -B, BEEx detected a high degree of batch effects among these cohorts. In case study MRI, BEEx revealed that case study MRI-B had a lower degree of batch effects than case study MRI-A. In case study CT, BEEx did not detect batch effects in the dataset.

BEEx reveals a high degree of batch effects in a WSI-based multicenter HCC dataset

In this case study, we examined a dataset that included four HCC WSI cohorts: GDPH, TCGA, FAHKU, and FAHZU. Clear batch effects were observed among the four cohorts in the dataset. However, BEEx did not identify any significant batch effects between GDPH and FAHKU.

As shown in **Fig. 4A**, differences in hematoxylin and eosin staining styles and intensities among the cohorts are noticeable. More specifically, TCGA and FAHZU exhibit significantly different staining styles from those of GDPH and FAHKU, whereas the staining styles of GDPH and FAHKU are not significantly different from each other. Similarly, FAHZU and TCGA each have distinct clusters, whereas GDPH and FAHKU overlap, as shown in **Fig. 4B**, indicating a high degree of batch effects among FAHZU, TCGA, and the combination of GDPH and FAHKU, as well as a low degree of batch effects between GDPH and FAHKU. Moreover, in the distribution analysis with violin plots, as shown in **Fig. 4C**, most tissue-, texture-, and color-based features (e.g., dark tissue ratio, stain H std, contrast RMS) distinguish all cohorts significantly ($P < 0.05$), whereas some features fail to distinguish cohorts (e.g., RGB 1 brightness mean) or can only separate one or two cohorts significantly (e.g., YUV 2 brightness mean and HSV 1 brightness std). Based on the results of hierarchical clustering (**Fig. 4D**), the samples from TCGA and FAHZU were primarily grouped based on their cohorts; however, the samples from GDPH and FAHKU were mixed. In the PVCA, as shown in **Fig. 4E**, the batch was the major contributor (63.48%) to the overall variation. Furthermore, the overall BES of case study WSI is 0.8417. The cohort-to-cohort BES values of TCGA to GDPH, TCGA to FAHKU, TCGA to FAHZU, GDPH to FAHKU, GDPH to FAHZU, and FAHKU to FAHZU are 0.9956, 0.9673, 0.9055, 0.5523, 0.9939, and 0.8839, respectively.

BEEx finds a low degree of batch effects in a TMA-based multicenter breast cancer dataset

In this dataset, all the TMAs were stained using the same reagents and antibodies and processed in three different laboratories. As shown in **Fig. 5A**, the variations in hematoxylin and eosin staining styles across these cohorts are barely discernible to the human eye, and the UMAP plot shown in **Fig. 5B** does not exhibit clear clusters. Nevertheless, a few subvisual features were detected by our distribution analysis violin plots (**Fig. 5C**), demonstrating that some color- and texture-based features can significantly differentiate these cohorts ($P < 0.05$, highlighted in red). This is because the staining of slides from different cohorts was conducted in different laboratories. In addition, the clustergram shown in **Fig. 5D** demonstrates that the samples are not grouped mainly based on the cohort to which they belong. However, most samples of VTMA-3 are clustered on the left and VTMA-2 on the right, reflecting batch effects between VTMA-3 and VTMA-2. As shown in **Fig. 5B**, VTMA-2 shares a different cluster below VTMA-1 and VTMA-3. The PVCA shown in **Fig. 5E** depicts the overall batch effects, in which batch contributes only 2.96% to the overall data variance, which is lower than overgrade (a tumor feature), which contributes 11.48%. Instead, the unknown random effect residuals explain most of the variance in the data. In addition, the overall BES of case study VTMA is 0.5072. The cohort-to-cohort BES values for VTMA-1 to VTMA-2, VTMA-1 to VTMA-3, and VTMA-2 to VTMA-3 are 0.5008, 0.5073, and 0.6339, respectively.

BEEx validates batch effect rectification with downstream task performance improvement in a MRI-based multicenter prostate cancer dataset

In this case study, we examined a dataset that consisted of three prostate cancer MRI cohorts: SSPH, RHWU, and YCPH. BEEx successfully revealed the existence of batch effects in case study MRI-A and -B and verified the reduction of batch effects in case study MRI-B, in which a batch effect rectification method was applied to case study MRI-A (10). In addition, the case with batch effect rectification (case study MRI-B) showed better downstream task performance than the case without batch effect rectification (case study MRI-A).

Here, we examine the batch effects of case study MRI-A, which did not apply a batch effect rectification method (10). **Figure 6A** shows that the image contrast intensities and resolutions for these cohorts differ. Noticeably, SSPH seems to have significantly smoother boundaries than the other two cohorts, possibly because the postprocessing algorithm was applied after image capture. In the UMAP visualization of the distribution analysis shown in **Fig. 6B**, three distinct clusters are evident according to

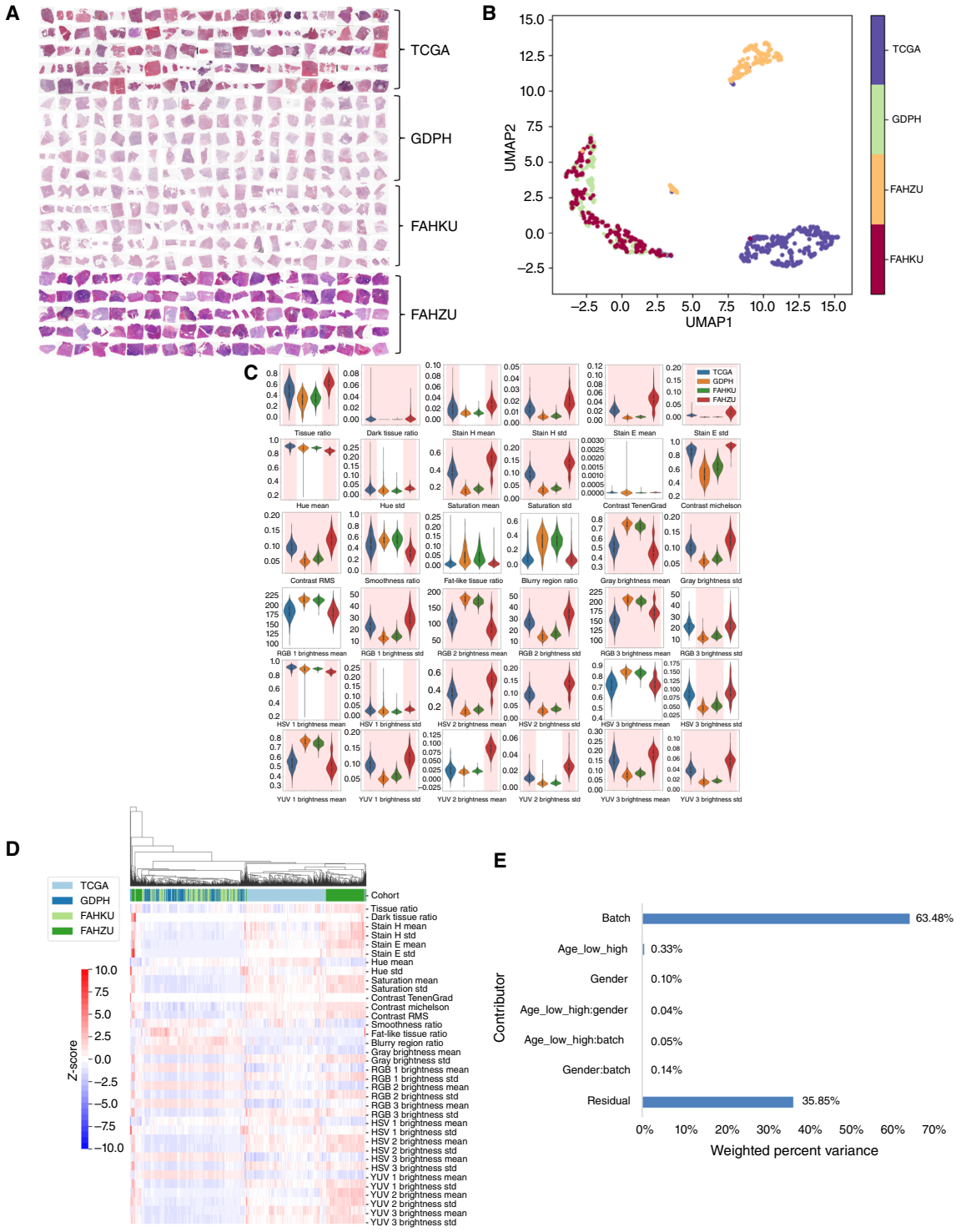

Figure 4.

Visualization of BEEx analyzers in case study WSI. In this case study, BEEx detected a high degree of batch effects. **A,** TCGA and FAHZU show a significantly different staining style from GDPH and FAHKU in image overviewer. Moreover, there is no apparent difference between the staining styles of GDPH and FAHKU. **B,** Similarly, in distribution analysis with UMAP, there are three noticeable distinct clusters, according to the cohorts. FAHZU and TCGA each have a separate cluster, whereas GDPH and FAHKU overlap, sharing the same separated cluster. **C,** In distribution analysis with violin plots, most features can separate these four cohorts significantly. **D,** In hierarchical clustering, almost all samples of TCGA and FAHZU are clustered to their cohort on the right, whereas the samples of GDPH and FAHKU are mixed on the left. **E,** In PVCA, batch explains 63.48%, making it the largest contributor to the overall data variance.

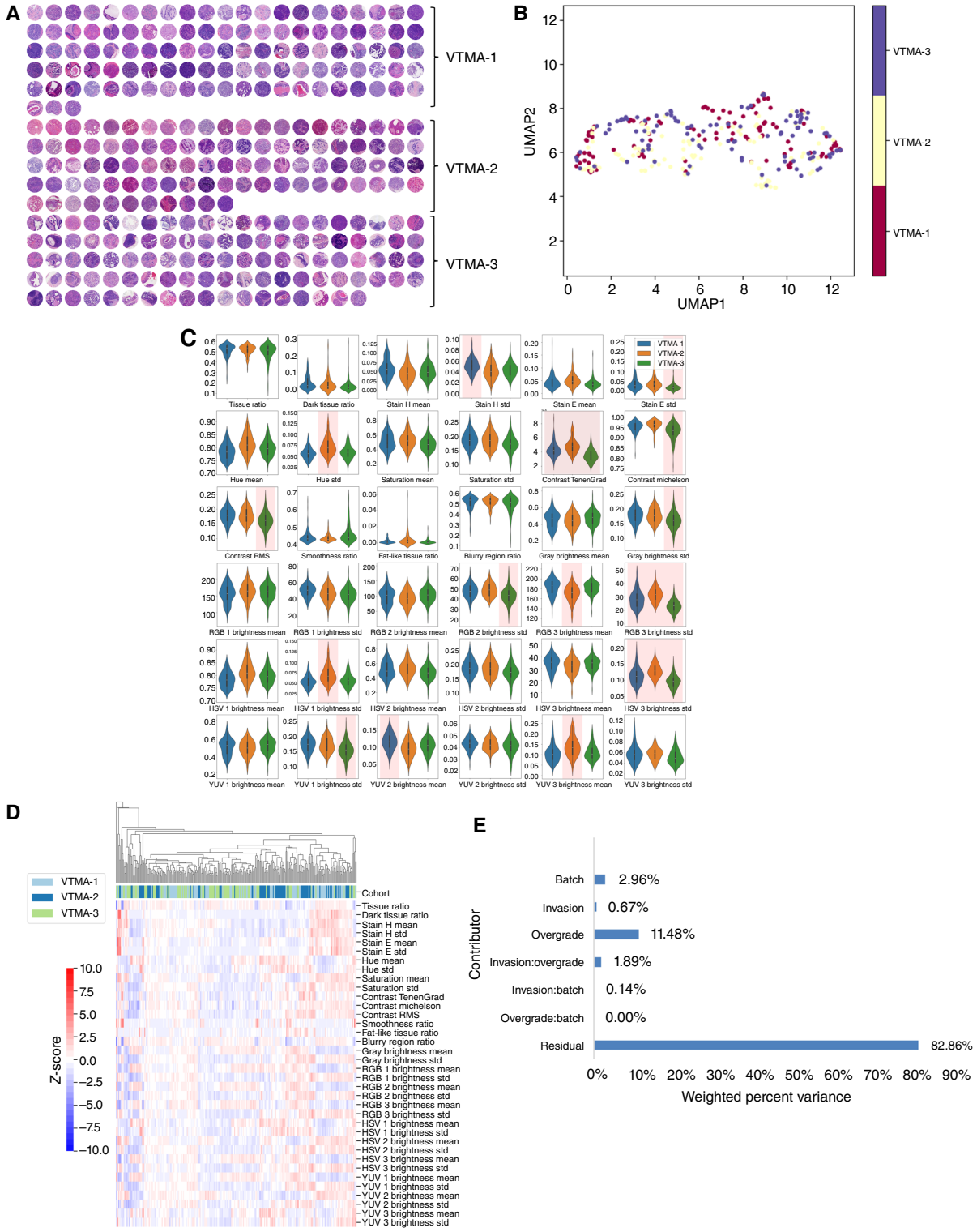

Figure 5.

Visualization of BEEx analyzers in case study VTMA. In this case study, BEEx demonstrates a low degree of batch effects. **A,** There are no clear color or shape differences in image overviewer. **B,** Moreover, distribution analysis with UMAP shows no distinct clusters, although VTMA-2 seems to share a different cluster below from VTMA-1 and VTMA-3. **C,** In distribution analysis with violin plots, several color- and texture-based features significantly separate these cohorts. **D,** In the hierarchical clustering, most samples of VTMA-3 are clustered on the left, whereas those of VTMA-2 are clustered on the right. **E,** Finally, in the PVCA, batch explains only 2.96% of the overall data variance.

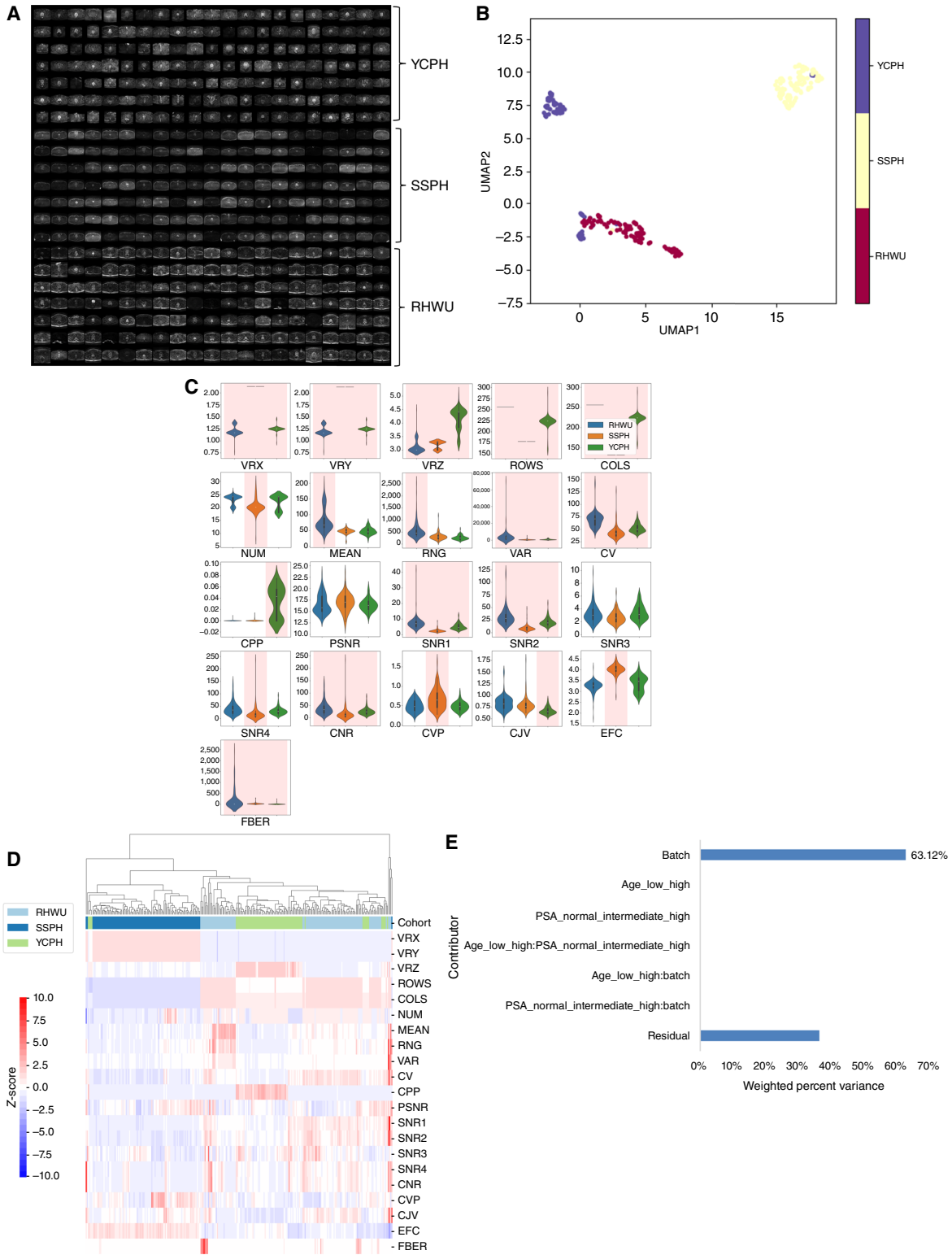

Figure 6.

Visualization of BEEx analyzers in case study MRI-A. **A,** Image overviewer shows the visual difference among the cohorts. **B,** At the same time, distribution analysis with UMAP shows three clearly distinct clusters, according to their cohorts. **C,** In distribution analysis with violin plots, most metadata and measurement features can significantly separate cohorts to some extent (with $P < 0.05$). **D,** In hierarchical clustering, almost all samples are grouped based on their cohorts. **E,** PVCA shows that batch explains 63.18%, making it the largest contributor to the overall data variance.

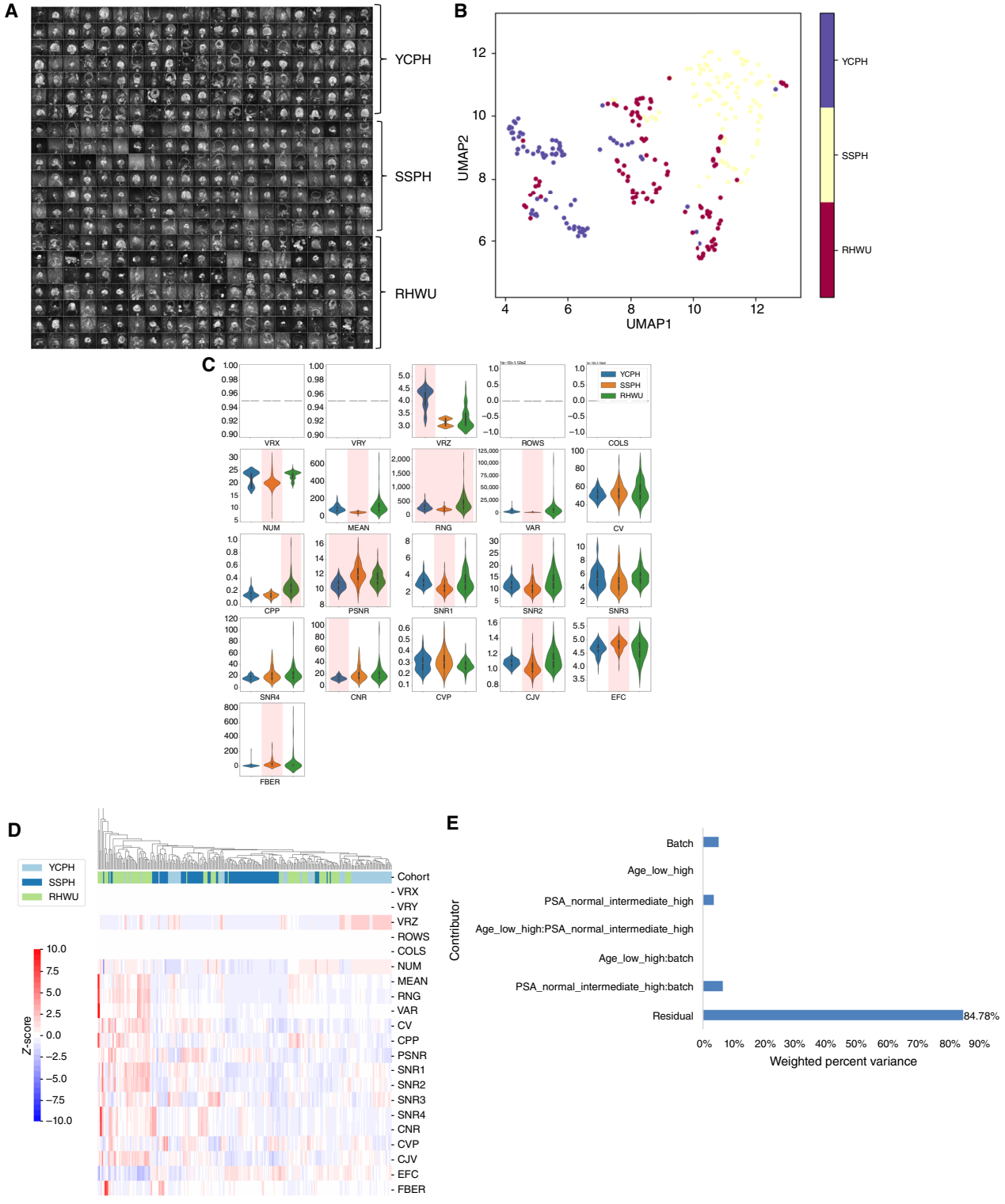

Figure 7.

Visualization of BEEx analyzers in case study MRI-B following the deployment of a batch effect rectification method, showing a decreased degree of batch effects compared with the results of case study MRI-A. **A,** Image overviewer shows the visual differences among cohorts. **B,** Distribution analysis with UMAP reveals three clusters based on the cohort, although some samples from RHWU overlapped with the other cohorts. **C,** Distribution analysis with violin plots demonstrates that most features cannot separate cohorts significantly or can only separate one cohort significantly. **D,** Hierarchical clustering shows that some samples are grouped according to cohorts, whereas other subgroups are mixed. **E,** The results of PVCA demonstrate that batch explains 5.08% of the overall data variance.

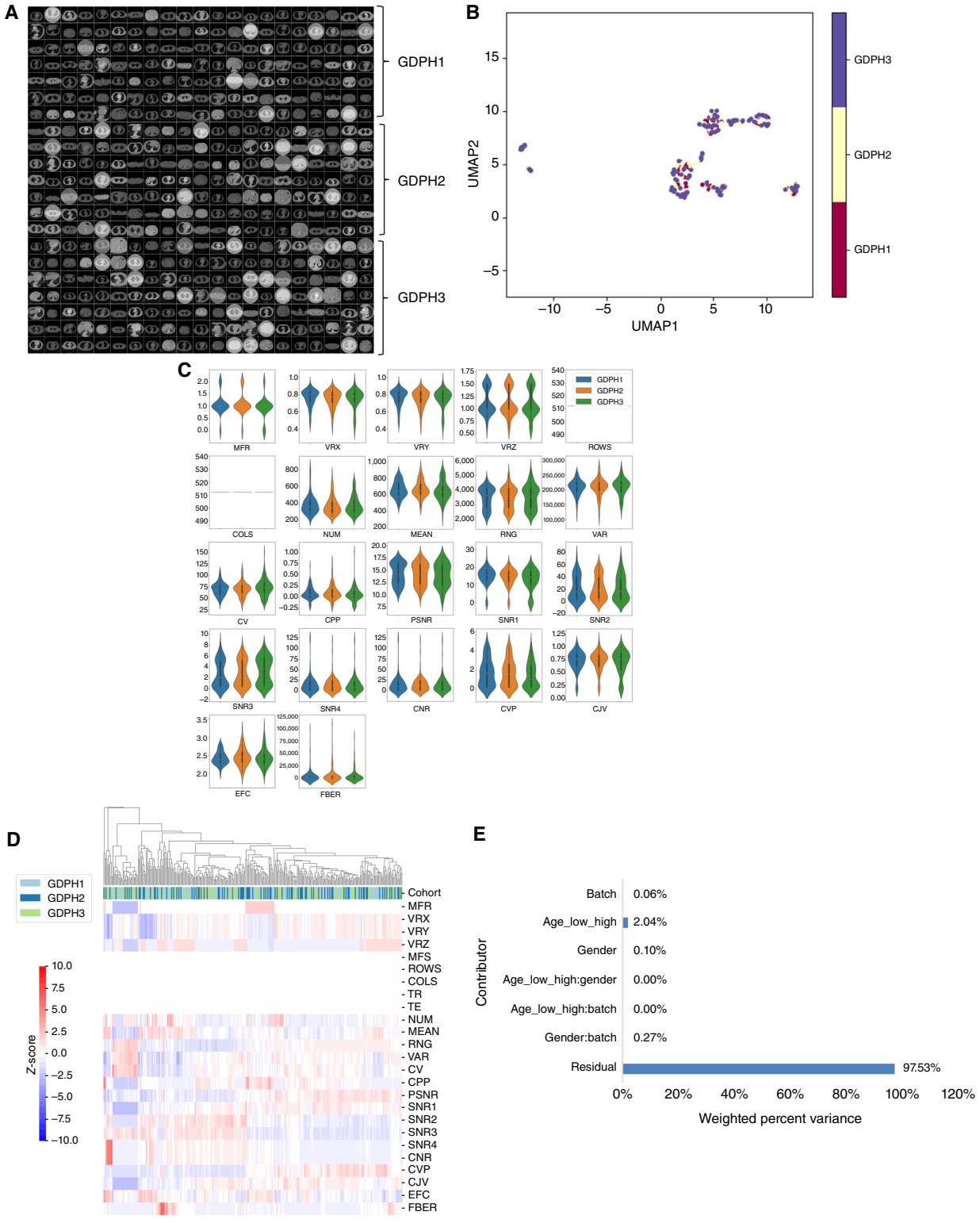

Figure 8.

BEEx results of Case Study-CT. This is a control case study that demonstrates no batch effects among all analyzers. **A,** There is no visual difference among these cohorts, as shown in image overviewer. **B,** In distribution analysis with UMAP, BEEx does not observe clearly separated clusters, according to the cohorts. **C,** In addition, distribution analysis violin plots shows that each cohort shares a similar feature distribution without significant differences. **D,** Similarly in hierarchical clustering, the samples are not grouped according to their cohort. **E,** Batch contributes only 0.06% to the overall variance in PVCA.

the cohorts. In the distribution analysis with violin plots shown in **Fig. 6C**, most metadata features (e.g., VRX and ROWS) distinguish these cohorts significantly ($P < 0.05$), and some measurement features can also separate cohorts (e.g., CPP and SNR1) significantly to some extent. In the clustergram shown in **Fig. 6D**, most samples are grouped based on their cohorts. In the PVCA shown in **Fig. 6E**, the batch variance (63.12%) is the primary contributor to the overall variation. The overall BES value of case study MRI-A is 0.8760. The cohort-to-cohort BES values of the RHWU to SSPH, RHWU to YCPH, and SSPH to YCPH are 0.9703, 0.9876, and 0.9554, respectively. In conclusion, BEEx found that case study MRI-A had a high degree of batch effects and that most features could significantly separate multicenter cohorts.

In case study MRI-B, we used a batch effect rectification method (10) that combined generative models (18) with an adversarial training strategy (19) to synthesize high–*b*-value MRI without rectal artifacts. After deploying batch effect rectification, the AUC performance for the downstream prostate cancer diagnosis task improved by 6% at the patient level (10). In parallel, BEEx successfully validated the decrease in batch effects in case study MRI-B compared with case study MRI-A. As shown in **Fig. 7A**, the image overviewer demonstrates less visual variation than that shown in **Fig. 6A**. Although the UMAP plot shown in **Fig. 7B** still has three clusters, some samples from different cohorts overlap, further indicating that feature-level differences are decreased via the batch effect rectification method. In the distribution analysis with violin plots, as shown in **Fig. 7C**, only two features (i.e., RNG and PSNR) can significantly ($P < 0.05$) separate three cohorts, whereas most features become unable to separate cohorts or can only separate one cohort. Similarly, in the clustergram shown in **Fig. 7D**, the result of sample grouping by cohorts is not as clear as before, as shown in **Fig. 6D**. In the PVCA in **Fig. 7E**, the data variance contributed by the batch decreases noticeably to 5.08%. In addition, the overall BES of case study MRI-B decreases to 0.7939. The cohort-to-cohort BES values of the RHWU to SSPH, RHWU to YCPH, and SSPH to YCPH were 0.7688, 0.7922, and 0.9662, respectively.

BEEx finds no batch effects in a CT-based control case study

Case Study-CT is a case–control study in which three pseudo cohorts, GDPH1, GDPH2, and GDPH3, were artificially created using randomly selected cases from the same dataset. As expected, BEEx found no batch effect among the cohorts. Specifically, there are no visual differences between the three cohorts (**Fig. 8A**). Moreover, the UMAP shown in **Fig. 8B** does not show clearly separated clusters, according to the cohorts. Similar to the violin plots shown in **Fig. 8C**, each subplot shares a similar distribution, indicating that cohorts GDPH1, GDPH2, and GDPH3 have the same distribution. In addition, the clustergram shown in **Fig. 8D** demonstrates that the samples are not grouped based on their cohorts. Finally, the batch contributes to 0.06% of the overall data variance, as shown in **Fig. 8E**. The overall BES value is 0.4987, and the cohort-to-cohort BES values of GDPH1 to GDPH2, GDPH1 to GDPH3, and GDPH2 to GDPH3 are 0.4976, 0.5009, and 0.5009, respectively, which are close to the results of random guessing.

Discussion

Recently, AI models have gained significant attention in addressing clinical challenges. However, owing to the heterogeneity of the tumor microenvironment, population variations, and interobserver variability among experts, researchers need to validate the reliability and applicability of extant AI models through multicenter studies (5, 20, 21). Nevertheless, batch effects, ubiquitous nonbiological variations in multicenter datasets, often affect model performance. A trained model affected by batch effects typically lacks generalizability and is prone to misinterpretations and wrong conclusions. Consequently, there is an urgent need for effective tools that can evaluate batch effects in multicenter settings.

To the best of our knowledge, existing platforms specifically designed to investigate batch effects neither support medical images (6) nor are they open-source (2, 3), as shown in Supplementary Table S3. Therefore, to empower researchers to conduct batch effects analysis prior to validating AI models in a multicenter study, we introduced BEEx, an open-source platform, for batch effect analysis. BEEx is a scalable and versatile framework designed to read, process, and analyze a wide range of medical images for batch effect evaluation. It is fully automated and features both a GUI and a command-line interface. Additionally, it includes configuration files that enable analysis across different cases and are freely available to the public. With the help of BEEx, users can quantify the severity of batch effects among cohorts, trace possible sources of batch effects, and gain insights to address and validate the methods used for batch effect rectification.

Focusing on pathologic images, we conducted two case studies in which BEEx effectively revealed batch effects via multiple visualization methods. In a case study involving radiologic image data, we utilized BEEx for batch effect examination before and after batch effect rectification and observed a significant reduction in the batch effect. Hence, we can use BEEx not only to observe batch effects in data prior to multicenter modeling but also as a validation tool for batch effect rectification methods.

The effectiveness of BEEx has been consistently demonstrated across multiple case studies, thereby reaffirming its utility in detecting and analyzing batch effects. These batch effects, which are often overlooked, can significantly affect the results of experiments by introducing artificial variation, generating spurious associations, reducing reproducibility, and resulting in misleading statistical analyses. Therefore, their identification and mitigation are of utmost importance in medical image analysis. As the features provided by BEEx are basic features used for image quality control, they may not fully meet the needs of all users. Nevertheless, because BEEx is an open-source framework, users are free to integrate additional features.

The ability of BEEx to detect these batch effects, as evidenced in case studies involving TMAs, WSIs, and MRI and CT images, shows its useful application in diverse research contexts. Researchers using BEEx as a prescreening tool prior to the development of a multicenter AI-based cancer diagnostic model can confidently assess batch effects in their data, identify their sources, and quantify their impact. This will allow them to mitigate batch effects by using appropriate normalization techniques to enhance the reliability and validity of their findings.

Authors' Disclosures

A. Madabhushi reports personal fees from Picture Health and SimBioSys, other support from Inspirata Inc. and Elucid Bioimaging, and grants from Bristol Myers Squibb and AstraZeneca during the conduct of the study, and other support from Takeda outside the submitted work. No disclosures were reported by the other authors.

Authors' Contributions

Y. Wu: Software, visualization, methodology, writing–original draft, writing–review and editing. **X. Xu:** Resources. **Y. Cheng:** Software. **X. Zhang:** Data curation. **F. Liu:** Methodology. **Z. Li:** Data curation. **L. Hu:** Data curation, funding acquisition. **A. Madabhushi:** Writing–review and editing. **P. Gao:** Data curation. **Z. Liu:** Data curation, supervision, funding acquisition. **C. Lu:** Resources, data curation, supervision, funding acquisition, validation, methodology, project administration, writing–review and editing.

Acknowledgments

This work was funded by The National Key R&D Program of China, 2023YFC3402800; National Natural Science Foundation of China (no. 82272084, no. 82302130); The National Natural Science Foundation of China Excellent Young Scientists Fund (Overseas; no. 22HAA01598); Regional Innovation and Development Joint Fund of National Natural Science Foundation of China (no. U22A20345); National Science Fund for Distinguished Young Scholars of China (no. 81925023); Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (no. 2022B1212010011); High-level Hospital Construction Project (no. DFJHBF202105); and Zhejiang Province Health Major Science and Technology Program of National Health Commission Scientific Research Fund (no. WKJ-ZJ-2426).

Note

Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Received February 12, 2024; revised July 11, 2024; accepted October 30, 2024; published first December 4, 2024.

References

1. Schmitt M, Maron RC, Hekler A, Stenzinger A, Hauschild A, Weichenthal M, et al. Hidden variables in deep learning digital pathology and their potential to cause batch effects: prediction model study. J Med Internet Res 2021;23:e23436.
2. Kothari S, Phan JH, Stokes TH, Osunkoya AO, Young AN, Wang MD. Removing batch effects from histopathological images for enhanced cancer diagnosis. IEEE J Biomed Health Inform 2014;18:765–72.
3. Stopsack KH, Tyekucheva S, Wang M, Gerke TA, Vaselkiv JB, Penney KL, et al. Extent, impact, and mitigation of batch effects in tumor biomarker studies using tissue microarrays. Elife 2021;10:e71265.
4. Wu Y, Koyuncu CF, Toro P, Corredor G, Feng Q, Buzzy C, et al. A machine learning model for separating epithelial and stromal regions in oral cavity squamous cell carcinomas using H&E-stained histology images: a multi-center, retrospective study. Oral Oncol 2022;131:105942.
5. Lu C, Bera K, Wang X, Prasanna P, Xu J, Janowczyk A, et al. A prognostic model for overall survival of patients with early-stage non-small cell lung cancer: a multicentre, retrospective study. Lancet Digit Health 2020;2:e594–606.
6. Zhu T, Sun R, Zhang F, Chen G-B, Yi X, Ruan G, et al. BatchServer: a web server for batch effect evaluation, visualization, and correction. J Proteome Res 2021;20:1079–86.
7. Sadri AR, Janowczyk A, Zhou R, Verma R, Beig N, Antunes J, et al. Technical Note: MRQy - an open-source tool for quality control of MR imaging data. Med Phys 2020;47:6029–38.
8. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: an open-source quality control tool for digital pathology slides. JCO Clin Cancer Inform 2019;3:1–7.
9. Chen Y, Zee J, Smith A, Jayapandian C, Hodgin J, Howell D, et al. Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. J Pathol 2021;253:268–78.
10. Hu L, Guo X, Zhou D, Wang Z, Dai L, Li L, et al. Development and validation of a deep learning model to reduce the interference of rectal artifacts in MRI-based prostate cancer diagnosis. Radiol Artif Intell 2024;6:e230362.
11. Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. Am Stat 1998;52:181–4.
12. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. JOSS 2018;3:861.
13. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. WIREs Data Mining Knowl Discov 2012;2:86–97.
14. Li J, Bushel PR, Chu T-M, Wolfinger RD. Principal variance components analysis: estimating batch effects in microarray gene expression data. In: Scherer A, editor. Batch effects and noise in microarray experiments. Chichester (UK): John Wiley & Sons, Ltd; 2009. p. 141--54.
15. Cuzick J. A Wilcoxon-type test for trend. Stat Med 1985;4:87–90.
16. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn 2003;52:91–118.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
18. Hu L, Zhou D, Zha Y, Li L, He H, Xu W, et al. Synthesizing high-b-value diffusion-weighted imaging of the prostate using generative adversarial networks. Radiol Artif Intell 2021;3:e200237.
19. Hu L, Zhou D, Xu J, Lu C, Han C, Shi Z, et al. Protecting prostate cancer classification from rectal artifacts via targeted adversarial training. IEEE J Biomed Health Inform 2024;28:1–14.
20. Zhao S, Chen D-P, Fu T, Yang J-C, Ma D, Zhu X-Z, et al. Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. Nat Commun 2023;14:6796.
21. Hölscher DL, Bouteldja N, Joodaki M, Russo ML, Lan Y-C, Sadr AV, et al. Next-generation morphometry for pathomics-data mining in histopathology. Nat Commun 2023;14:470.