Where is VALDO? VAscular Lesions Detection and segmentatiOn challenge at MICCAI 2021

Carole H. Sudre[a,b,c,1,*], Kimberlin Van Wijnen[d,*], Florian Dubost[d], Hieab Adams[e], David Atkinson[f], Frederik Barkhof[b,g], Mahlet A. Birhanu[d], Esther E. Bron[d], Robin Camarasa[d], Nish Chaturvedi[a], Yuan Chen[i], Zihao Chen[j], Shuai Chen[d], Qi Dou[k], Tavia Evans[e], Ivan Ezhov[l,x], Haojun Gao[m], Marta Girones Sanguesa[n], Juan Domingo Gispert[o,ab,ac], Beatriz Gomez Anson[p], Alun D. Hughes[a], M. Arfan Ikram[q], Silvia Ingala[g], H. Rolf Jaeger[r], Florian Kofler[l,w,x], Hugo J. Kuijf[n], Denis Kutnar[n], Minho Lee, Bo Li[d], Luigi Lorenzini[g], Bjoern Menze[l,y], Jose Luis Molinuevo[o,aa], Yiwei Pan[s], Elodie Puybareau[t], Rafael Rehwald[r], Ruisheng Su[d], Pengcheng Shi[s], Lorna Smith, Therese Tillin[a], Guillaume Tochon[t], Hélène Urien[u], Bas H.M. van der Velden[n], Isabelle F. van der Velpen[h,q], Benedikt Wiestler[w], Frank J. Wolters[h,q], Pinar Yilmaz[q], Marius de Groot[d,z], Meike W. Vernooij[h,q], Marleen de Bruijne[d,v], for the ALFA study

[a] *MRC Unit for Lifelong Health and Ageing at UCL, University College London, London, United Kingdom*
[b] *Centre for Medical Image Computing, University College London, London, United Kingdom*
[c] *School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom*
[d] *Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
[e] *Department of Clinical Genetics and Radiology, Erasmus MC, Rotterdam, The Netherlands*
[f] *Centre for Medical Imaging, University College London, London United Kingdom*
[g] *Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centre, Amsterdam, The Netherlands*
[h] *Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
[i] *Department of Radiology, University of Massachusetts Medical School, Worcester, The USA*
[j] *School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China*
[k] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, China*
[l] *Department of Informatics, Technische Universitat Munchen, Munich, Germany*
[m] *Department of Radiology, Zhejiang University, Hangzhou, China*
[n] *Image Sciences Institute, University Medical Center Utrecht, Utrecht, the Netherlands*
[o] *Barcelonaβ Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain*
[p] *Department of Radiology, Hospital San Pau i santa Creu, Barcelona, Spain*
[q] *Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands*
[r] *Institute of Neurology, University College London, London, United Kingdom*
[s] *Department of Electronic and Information Engineering, Harbin Institute of Technology at Shenzhen, Shenzhen, China*
[t] *LRDE, EPITA, Paris, France*
[u] *ISEP-Institut Supérieur d'Électronique de Paris, Issy-les-Moulineaux, France*
[v] *Department of Computer Science, University of Copenhagen, Copenhagen, Denmark*
[w] *Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Germany*
[x] *TranslaTUM - Central Institute for Translational Cancer Research, Technical University of Munich, Germany*
[y] *Department of Quantitative Biomedicine, University of Zurich, Switzerland*
[z] *GlaxoSmithKline Research, Stevenage,UK*
[aa] *H. Lundbeck A/S, Copenhagen, Denmark*
[ab] *Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain*
[ac] *Centro de Investigación Biomédica en Red Bioingeniería, Biomateriales y Nanomedicina, (CIBER-BBN), Barcelona, Spain*

*Authors contributed equally to this work
[1]Corresponding author:c.sudre@ucl.ac.uk
[2]The complete list of collaborators for the ALFA study can be found in acknowledgments

Abstract

Imaging markers of cerebral small vessel disease provide valuable information on brain health, but their manual assessment is time-consuming and hampered by substantial intra- and interrater variability. Automated rating may benefit biomedical research, as well as clinical assessment, but diagnostic reliability of existing algorithms is unknown. Here, we present the results of the *VAscular Lesions DetectiOn and Segmentation* (*Where is VALDO?*) challenge that was run as a satellite event at the international conference on Medical Image Computing and Computer Aided Intervention (MICCAI) 2021. This challenge aimed to promote the development of methods for automated detection and segmentation of small and sparse imaging markers of cerebral small vessel disease, namely enlarged perivascular spaces (EPVS) (Task 1), cerebral microbleeds (Task 2) and lacunes of presumed vascular origin (Task 3) while leveraging weak and noisy labels. Overall, 12 teams participated in the challenge proposing solutions for one or more tasks (4 for Task 1 - EPVS, 9 for Task 2 - Microbleeds and 6 for Task 3 - Lacunes). Multi-cohort data was used in both training and evaluation. Results showed a large variability in performance both across teams and across tasks, with promising results notably for Task 1 - EPVS and Task 2 - Microbleeds and not practically useful results yet for Task 3 - Lacunes. It also highlighted the performance inconsistency across cases that may deter use at an individual level, while still proving useful at a population level.

Keywords: CSVD, brain, MRI, microbleeds, enlarged perivascular spaces, lacunes, automated, segmentation, detection, challenge

1. Introduction

Cerebral small vessel disease (CSVD), the deterioration of the smallest brain vessels, encompasses a large variety of etiologies including arteriolosclerosis (Alistair, 2002) and amyloid pathology (Kester et al., 2014) and may be further driven by genetic predisposition (Haffner et al., 2016; Giau et al., 2019). It results in observable damage or changes to the brain. Most commonly observed MRI markers of CSVD include white matter hyperintensities (WMH), cerebral microbleeds, lacunes of presumed vascular origin, and enlarged perivascular spaces (Wardlaw et al., 2013). CSVD related damage has been associated with an increased risk of stroke and dementia, and with the acceleration of cognitive decline (Østergaard et al., 2016; Rensma et al., 2018). The presence of

these markers are also associated to one another (Zhang et al., 2014; Zhu et al., 2010; Yates et al., 2014).

WMH are the most visible marker of CSVD and have naturally taken the centre stage of clinical research in CSVD. In addition, research on development of WMH segmentation solutions has been particularly popularized thanks to impactful research showing the clinical importance of lesion volumetry (Van Straaten et al., 2006). While the automated quantification of white matter hyperintensities has been heavily studied for the last decade with very successful solutions (Sudre et al., 2015; Guerrero et al., 2018; Atlason et al., 2019; De Boer et al., 2009), automated detection and segmentation of the small, focal markers of CSVD has been investigated less frequently. However, as the interest of the clinical community in these markers starts to grow, getting to understand their relevance in clinical research requires them to be adequately detected and quantified. While these markers are currently typically assessed visually through binary dichotomization (presence vs absence) (Yates et al., 2014), counts (Adams et al., 2015), or visual scales (Potter, 2011), such visual assessment is time consuming and subject to large inter- and intra-rater variability (Sudre et al., 2019). Automated methods are therefore required to make quantification robust and reliable as well as feasible in the context of large data sets. So far, development of automated methods has been impeded by the methodological issues related to the very small size of these markers and the resulting extreme imbalance in the data, as well as the absence of a gold standard for annotation.

Methodological developments towards automated solutions for the quantification of biomarkers have found a new dynamic thanks to the annotated datasets made available through technical challenges on segmentation and detection in brain MRI with notably the popular BRATS challenge (Menze et al., 2014), ISLes (Maier et al., 2017), MRBrainS (Mendrik et al., 2015), the 2017 MICCAI WMH challenge (Kuijf et al., 2019) or the more recent ADAM challenge (Timmins et al., 2021) on intracranial aneurysms. Such challenges give insight into state-of-the-art methodology and remaining technical problems for a specific question.

The *VAscular Lesions DetectiOn and Segmentation* (*Where is VALDO?*) challenge was organized with the aim of promoting the development of new solutions for the automated detection and segmentation of these sparse and small structural brain changes (enlarged perivascular spaces (Task 1), cerebral microbleeds (Task 2) and lacunes (Task 3)) while leveraging weak and noisy labels from manual annotation or visual assessment. Beyond a simple benchmarking exercise assessing the state of the solution space, this challenge was further intended to gain insight on the current

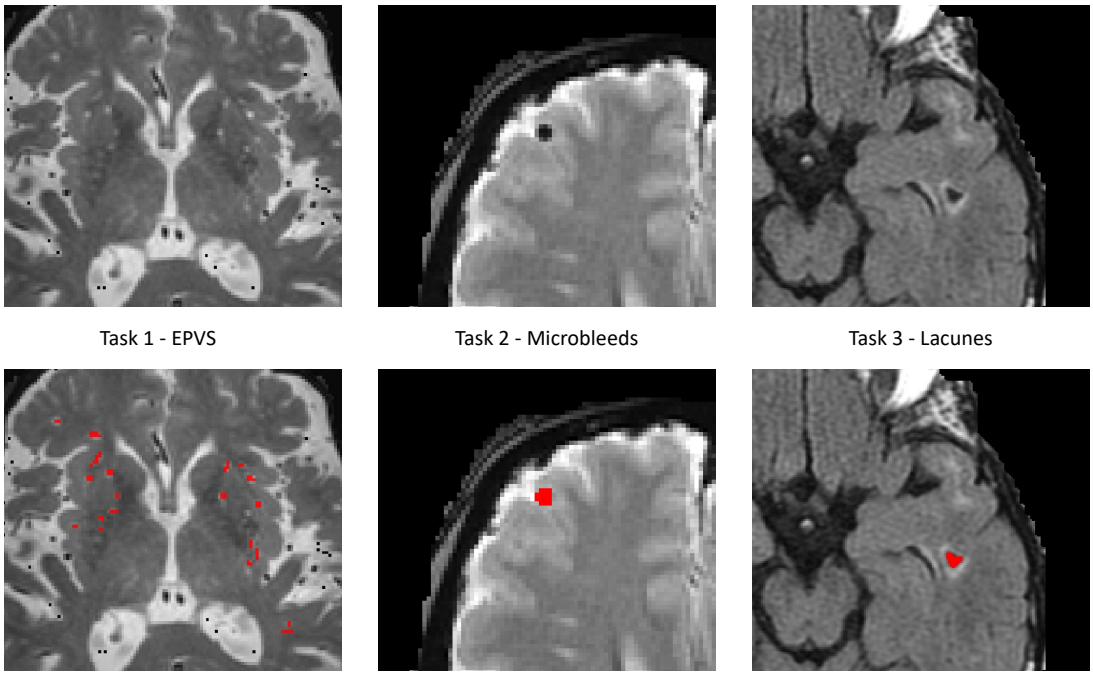

Figure 1: Annotated example of the three type of markers targeted in the challenge

pitfalls and challenges, raise awareness and contribute to the building of a community dedicated to developing solutions to facilitate quantification of CSVD markers in brain MRI scans. This paper describes the design, results, and lessons learnt through the challenge according to the reporting guidelines detailed in (Maier-Hein et al., 2020).

2. Methods

2.1. Mission of the challenge

The *Where is VALDO?* challenge was organized to assess three tasks, each of them focusing on one focal marker of CSVD - Task 1 on enlarged perivascular spaces (EPVS), Task 2 on cerebral microbleeds and Task 3 on Lacunes. Figure 1 illustrates each of these markers as annotated in the challenge training set.

Currently, the lack of accurate and reproducible automated methods for all three markers prohibits the identification of clinically relevant characteristics at both individual and population levels. Therefore, for each of the stated markers both detection and segmentation performance need to be assessed. Ultimately, the improved quantification of these small focal markers of CSVD may be used to better understand their relevance and derive biomarkers for diagnosis or prognosis in the context of healthy ageing and dementia, and as surrogate end points in clinical trials.

In proposing tasks particularly subject to high data imbalance and limited and/or noisy annotations, this challenge further aimed to catalyse methodological research to address these common issues in the medical image analysis community.

Ultimately, the proposed methods should be applicable to different settings involving ageing populations such as population cohorts, clinical trials or memory clinics.

The challenge dataset however consisted exclusively of population-based cohorts - two to three according to the task, with differences in MRI acquisition protocol, image resolution and scanner characteristics across datasets. No additional information beyond the images was provided. Each of the datasets was enriched for lesion burden through stratified sampling of the skewed population distributions.

For each task, a similar approach to assessment was adopted to ensure consistency across tasks and address both segmentation and detection aspects, although some may currently be considered more important in one task than another, with different paradigms used in clinical practice. For instance, the blooming effect observed in the presence of microbleeds is protocol dependent, making the detection more relevant than the segmentation in that task (Buch et al., 2017).

2.2. Challenge organization

The *Where is Valdo?* challenge was run as a satellite event at MICCAI 2021 as a collaboration of University College London and Erasmus MC University Medical Center Rotterdam. Its three-task design was peer-reviewed prior to acceptance and made public at `https://doi.org/10.5281/zenodo.4600654` Regarding prize eligibility, it was decided that organizers would not participate and while members of the same institutions as the organizers were allowed to participate in the challenge, they would not be eligible for prizes. Prizes were given to each winner of individual tasks and the overall winner across all tasks. Results were publicly presented for all participating teams. All submitting teams were invited to propose two team members (per task) to participate

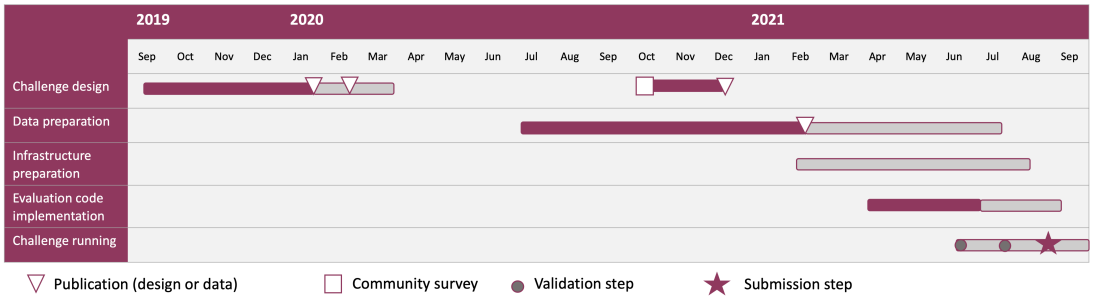

Figure 2: Timeline of the challenge from inception in September 2019

as co-authors in the challenge overview paper. After publication of this overview paper of the challenge, the submission will reopen to the community for anyone wanting to benchmark their methods against those previously submitted. Further information is available on the challenge website `https://valdo.grand-challenge.org`.

The challenge was organized in 4 phases: 1) a training phase from the moment the annotated database was made downloadable (February 2021), 2&3) two optional validation steps on 5 new cases to provide individual (no public leaderboard) feedback on the performance (14^{th} to 21^{st} of June and 12^{th} to 19^{th} of July) and 4) the final evaluation stage on withheld cases (submission from 26^{th} of July to 5^{th} of August 2021). A grace period extending until the 10^{th} August in case of technical difficulties was granted to all participants. Participants had to provide a docker container for their fully automated method (1 for each task) and were allowed to participate in any or all the tasks. Use of additional training data was allowed under the condition it would be made available at submission time. The methods did not have to be similar across all tasks. Details of the submission procedure are listed at `https://valdo.grand-challenge.org/Submission/`. Participating teams were also requested to provide a short technical note describing their solutions that have been made available at `https://openreview.net/group?id=MICCAI.org/2021/Challenge/VALDO`. Figure 2 presents the timeline of the challenge.

Submitted data were evaluated on the test set at a GPU facility at Erasmus MC. In order to ensure that the proposed methods were running as expected, each docker was run on one example of the training set and the result sent back to the participants for checking, allowing for submission of a new docker if the output was not as expected.

The evaluation code was made available prior to submission at `https://github.com/WhereIsValdo/`

`valdo-eval-2021`. The participating teams were encouraged to make their source code publicly available and all participants except one team agreed for their docker containers to be made public. They have been placed on `https://hub.docker.com/r/whereisvaldo/challenge2021/tags`

The challenge was sponsored by NVIDIA and Icometrix. Test data was available to CHS and KVW. The contribution of the authors listed in this manuscript can be found in supplementary material.

2.3. Community survey

To better understand the interest within the community for such initiative, we launched in January 2021 a survey targeting the community working in the field of automated detection of CSVD lesions. This survey was sent to a list of researchers having recently published automated methods for detection or segmentation of one of the three lesion types considered in the *Where is VALDO?* challenge, the International Society of Vascular Behavioural and Cognitive Disorders (VasCog `https://www.vas-cog.com`), and the Medical Image Understanding and Analysis (MIUA `miua@jiscmail.ac.uk`) mailing list, and the survey was shared on social media by the challenge organizing team. Overall, 36 answers were recorded with 25 individuals indicating to be very likely or likely to participate. Among the respondents, 39% indicated being already actively working in the field of CSVD and 30% more general in the neuroimaging field. Microbleed segmentation appeared as the most popular task in the survey with 15 respondents indicating they were highly likely to participate in this task against 10 for EPVS and 10 for lacunes. These answers helped shape the final challenge design, notably standardizing the evaluation of the different tasks and making the challenge overall more concise.

2.4. Challenge data sets

The challenge data sets (training, validation, and test sets) came from the same cohorts with a similar ratio between them across tasks. This ratio was also kept in the testing set.

2.4.1. Datasets and image acquisition

Two subsets of population cohorts were used for all three tasks and an additional one was further available for the microbleed detection/segmentation task, namely the SABRE and Rotterdam Scan Study (RSS) cohorts and the ALFA study respectively. All cohorts were retrospective studies for which local ethical approval had already been obtained from the National Research Ethics Service

Committee, London-Fulham (14/LO/0108) for SABRE, the Population Research Act from the Ministry of Health for RSS and the Independent Ethics Committee Parc de Salut Mar Barcelona and registered at Clinicaltrials.gov (NCT01835717) for ALFA. For all datasets, acquisition of the data was performed by a trained radiographer according to a predefined research protocol. The training data for the *Where is VALDO?* challenge was made available under a CC BY NC-SA license.

SABRE. The Southall and Brent Revisited (SABRE) cohort is a population cohort of individuals residing in the two named boroughs of west London (UK)(Tillin et al., 2013). This tri-ethnic cohort was initially recruited in 1988 with the purpose of investigating metabolic and cardiovascular diseases across ethnicities. For their third clinical visit (2014-2018), life partners were also invited to take part and study participants underwent a brain MRI session on a Philips 3T scanner. Mean age in this cohort at time of acquisition was 72 years old ranging from 36 to 92.

RSS. The Rotterdam Scan Study (RSS) (Ikram et al., 2015) is part of the larger Rotterdam Study (RS) (Ikram et al., 2020), a population-based study that aims to investigate chronic illness in the elderly. Started in 1995, the Rotterdam Scan Study initially concerned a selection of the RS but since 2005 brain MRI is part of the core protocol of the study. Individuals aged 45 and over without dementia are eligible for MRI and are followed up every 3-4 years. Since 2005, scanning has been performed on a 1.5T GE MRI scanner dedicated to the study.

ALFA. The ALFA (Alzheimer's and Families) cohort is based on the ALFA registry that gathers details of relatives (generally offspring) of patients with Alzheimer's Disease making up for a cohort naturally enriched for genetic predisposition to AD. As described in the related protocol paper (Molinuevo et al., 2016), the ALFA cohort is composed of cognitively normal participants aged 45-74. Brain MRI sequences were acquired on a GE Discovery 3T scanner.

Table 1 summarizes the acquisition parameters for the different sequences across the studied cohorts.

2.4.2. Training, validation and testing data

For Task 1 - EPVS and Task 3 - Lacunes, imaging data consists of T1-weighted, T2-weighted and FLAIR images, with the latter two modalities rigidly registered to the T1 image using NiftyReg (Modat et al., 2014). For Task 2 - Microbleeds, imaging data is the combination of T2, T2* and

Cohort	Sequence	Type	TR	TE	TI	FA	Resolution (mm)
SABRE	T1w	Inversion prepared gradient echo	6.9	3.1	/	/	1.09 x 1.09 x 1.0
	T2w	3D sagittal turbo spin echo	2500	222	836	8	1.09 x 1.09 x 1.0
	FLAIR		4800	125	1650		1.09 x 1.09 x 1.0
	T2*	Gradient echo	1288	21	/	18	0.45 x 0.45 x 3.0
RSS	T1w	Gradient recalled echo	13.8	2.8	400	20	0.49 x 0.49 x 0.8
	T2w	Fast spin echo	12300	17.3	/	/	0.49 x 0.49 x 0.8
	FLAIR	Fast spin echo	8000	120	2000		0.49 x 0.49 x 0.8
	T2*	Gradient recalled echo	45	31	/	13	0.49 x 0.49 x 0.8
ALFA	T1w	3D	8.0	3.7	450	8	1.0 x 1.0 x 1.0
	T2w	Fast spin echo	5000	85	/	110	1.0 x 1.0 x 3.0
	T2*	Gradient recalled echo	1300	23	/	15	1.0 x 1.0 x 3.0

Table 1: Acquisition details for the three cohorts. Acronyms FA - Flip angle; TE - echo time(ms); TI - inversion time(ms); TR - repetition time (ms)

	Task 1 - EPVS		Task 2 - Microbleeds		Task 3 - Lacunes	
Cohort	Train	Test	Train	Test	Train	Test
SABRE	6	10	11	20	6	10
RSS	34 (6/28)	56	34	68	34	56
ALFA	/	/	27	59	/	/
Total	40	66	72	147	40	66

Table 2: Number of cases in train and test set for each task and cohort origin. For RSS Task 1 of training separation between cases with full annotation and cases with only counts

T1-weighted images in T2* space. Table 2 presents the number of cases used for training and testing across the different tasks and the different cohorts. For each task, validation consisted of 5 cases from the RSS cohort. There was no overlap between training, test or validation datasets.

The number of cases proposed for training was chosen based on annotation availability and data policy for making a certain number of cases publicly available. For Task 1 - EPVS and Task 3 - Lacunes, the SABRE segmentation data was already available for a set of 16 cases with high level of cerebrovascular damage. In comparison, for the RSS study, for which annotations were more widely available, data were selected to cover the variability in burden present in the study. They present close to a uniform distribution in burden thereby limiting data skewness towards cases without any lesion. In all tasks, annotated cases were distributed across training and testing set to follow approximately similar burden distribution. A ratio of 6:10 between training and testing data was chosen across all cohorts and tasks.

2.4.3. Annotation

Across the three cohorts, raters were all trained for their annotation task and had at least 3 years of professional experience in dealing with medical images. The segmentation was performed for all SABRE and ALFA cases using ITKSnap (Yushkevich et al., 2016). For the RSS cases a custom MeVisLab (Ritter et al., 2011) application was used. In all cases were two annotations were available, the average of the two annotations was used as reference.

Task 1 - Enlarged Perivascular Spaces. For Task 1, the annotation strategy differed between the SABRE and RSS cohort. For identifying EPVS, the STRIVE criteria (Wardlaw et al., 2013) for EPVS were used in the SABRE cohort, while in the RSS cohort, the UNIVRSE criteria (Adams

et al., 2015) were used. These criteria are very similar, except for the fact that the UNIVRSE criteria only consider EPVS with a diameter between 1 and 3 mm, while the STRIVE criteria do not have a lower limit and consider any EPVS with a diameter up to 3 mm. In the SABRE cohort, EPVS over the whole brain image were annotated independently by two raters (CHS and LL) with a senior radiologist (BGA) confirming the segmentation of CHS. The three modalities were jointly used for the segmentation that was assessed across the three axes.

For this dataset the annotation was provided in either of two forms: over the full brain or on only 5 randomly selected slabs of 5mm. A mask was provided per case indicating the slabs that were annotated.

In the RSS cohort, EPVS were annotated with segmentations in limited axial slices for 6 cases of the training set and the full test set, while the remaining 28 cases of the training set were annotated with dots only by a team of trained annotators supervized by KVW, FD and MWV. EPVS were annotated in four brain regions: the mesencephalon, hippocampus, basal ganglia, and the centrum semi-ovale. The first two smaller regions were annotated entirely. For the latter two regions, only one fixed slice was annotated. For the cases with EPVS segmentations, additional slices of the basal ganglia and of the white matter were annotated, the depth of these axial slices was randomly chosen per case. A mask indicating which parts of the brain had been annotated was computed using parcellation outputs for each case.

For the training data made available to participants, the EPVS annotations were either presented just as counts (computed from the dots), per slice and per region or as segmentations plus counts in the same areas. The masks indicating the annotated regions and slices per case was also provided. Figure 3 illustrates the type of annotation masks that were provided to the participants.

Task 2 - Microbleeds. Different raters annotated each of the cohorts but followed very similar protocols. The BOMBS criteria (Cordonnier et al., 2009) was applied for the SABRE (RR under the supervision of HRJ) and ALFA cohort (consensus of SI and LL under the supervision of FB) as described in (Ingala et al., 2020). A team of trained raters under the supervision of MWV applied the protocol described in (Vernooij et al., 2008) for RSS. Both identification protocols are in line with the STRIVE guidelines (Wardlaw et al., 2013) that indicate that microbleeds are areas of signal void of generally 2-5 mm in diameter but can be up to 10 mm.

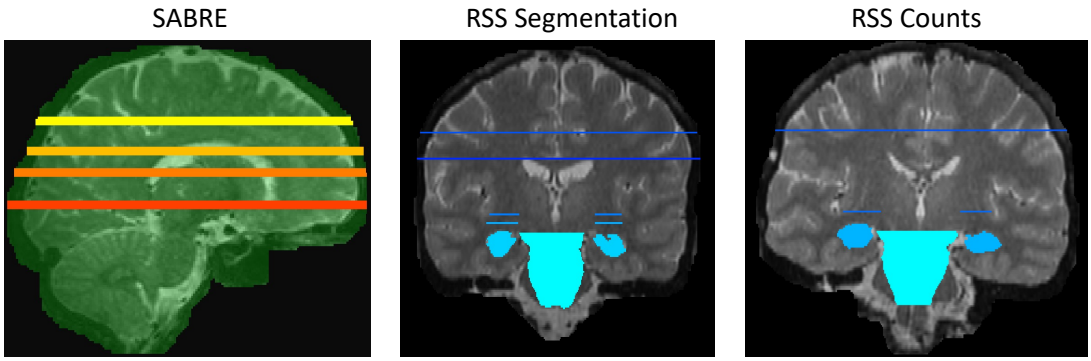

Figure 3: Example of annotation provided for Task 1 - EPVS with left) for SABRE slabs of 5 mm randomly selected or full segmentation over the image, middle) Segmentation on two slices of CSO, 2 slices of the basal ganglia, the hippocampi and mesencephalon for 6 RSS cases and right) count of EPVS on 1 slice of CSO, 1 slice of basal ganglia, hippocampi and mesencephalon for 28 cases of RSS.

Task 3 - Lacunes. Lacunes were identified using the STRIVE criteria (Wardlaw et al., 2013). Cerebellar lacunes were excluded because of assumed differences in the underlying pathology in this brain region(Sigurdsson et al., 2022). Any surrounding gliosis (the hyperintense rim visible on FLAIR sequences) was not included in the segmentation of the lacune. For the SABRE cohort, lacunes were identified at the same time as EPVS simply being assigned another label in the segmentation, with the two raters (CHS, LL) performing the identification and segmentation independently. For the RSS cohort, lacunes were independently segmented for all cases by two raters, the pair of raters varying across the cases. In RSS, all cases of training, validation and test set indicated by radiological reads as containing at least one lacune were consistently annotated by one rater (TE) on a custom MeVisLab(Ritter et al., 2011) application. The second set of annotations was performed using ITKSnap(Yushkevich et al., 2016). PY annotated all cases of the training set. FW annotated the validation set as well as half of the test set. The remaining half of the test set was annotated by IFV.

2.4.4. Sources of annotation errors

In all tasks, possible source of errors in the annotations pertain to multiple distinct sources: the appropriate identification of a target element either because these elements are very small and may be easy to miss or because it may be difficult to distinguish them from similarly appearing structures (mimics); the decision on the boundary of an object, probably notably more complex in

a coarser resolution plane; the use of the segmentation software (too large brush, not considering all orientations for consistency or not adequately using the zoom). In the case of EPVS, identification of "large enough" marker was also a subjective consideration possibly leading to different detection levels.

2.4.5. Preprocessing

For all tasks, the preprocessing consisted of a rigid alignment of the images as indicated in section 2.4.2. A defacing mask derived from the T1-weighted image was applied to all registered modalities. While such a step would not be required in practice, this step was mandated by the data sharing policies around public release of the data. The defacing mask was obtained as the inverse of a dilated version of the brain mask as obtained from HD-BET (Isensee et al., 2019). All RSS scans were corrected for intensity inhomogeneity with the default parameters of MINC N3 package (Sled et al., 1998).

2.5. Assessment method

All three tasks were evaluated using similar metrics in order to assess both detection and segmentation performance of the proposed solutions. A combination of relative error (F1 score and Mean Dice score) and absolute error (absolute element difference (AED) and absolute volume difference (AVD)) metrics was chosen, since they provide complementary information. The F1 score and the AED on the number of detected lesions were chosen as detection metrics while the Mean Dice score over the appropriately identified elements and the AVD were the metrics used for the evaluation of segmentation. Table 3 summarizes the purpose, formula and properties of the metrics used in the challenge across all tasks and calculated for each case, where $_c$ refers to 6-neighborhood connected components, TP to true positives, FP to False positive, FN to false negatives, Ref to the reference annotation and Seg to the predicted segmentation.

One essential aspect in the evaluation for the derivation of both F1 and Mean Dice score was the definition of true positive elements. To determine which of the elements were true positives, for all three tasks, connected components with a neighbourhood of 6 were established for both annotation and prediction using a threshold for the probability of 0.5 for the prediction map. Each annotation element was matched to at most one element from the prediction. For Task 1 - EPVS, a possible matchable element had to have an Intersection over Union (IoU) of more than 10%. For Task 2 - Microbleeds and Task 3 - Lacunes, matching was possible when the centre of mass

Metric	Target	Formula	Range	Best
F1 Score	Detection	$\frac{2TP_c}{2TP_c+FP_c+FN_c} * 100$	0 - 100	100
AED	Detection	$\lvert \#_c Ref - \#_c Seg \rvert$	0 - inf	0
Mean Dice	Segmentation	$\frac{100}{\#TP_c} \sum_{t \in TP_c} \frac{2 * \sum (Ref_t * Seg_t)}{\sum Ref_t + \sum Seg_t}$	0 - 100	100
AVD	Segmentation	$\lvert Ref - Seg \rvert$	0 - inf	0

Table 3: Description of detection and segmentation metrics used across all tasks for the evaluation.

of the prediction element was less than 5 mm away from the center of mass of the ground truth segmentation element. When multiple elements were found to be matchable, the one with best association value (IoU or centre of mass distance) was attributed to the annotated label. For empty cases, the relative metrics were inapplicable, so only the absolute error metrics (number of elements and volume) were computed.

In the event of algorithmic failure for a specific case, worst metric values were attributed. For bounded metrics (F1 and Mean Dice score) a value of 0 was given. For non-bounded error metrics (absolute element and absolute volume difference) an error of 100 000 was assigned as worst possible error.

	Detection Error	Detection Error	True Positive	True Positive
	RefUnc <= 0.5	RefUnc > 0.5	RefUnc <=0.5	RefUnc > 0.5
PredUnc<= 0.5	FC	FC	TC	FC
PredUnc >0.5	TU	TU	FU	TC

Table 4: Categorization for calculation of uncertainty measures; TU - Truly Uncertain; TC - Truly Certain; FU - Falsely Uncertain; FC - Falsely Certain

For Task 3 - Lacunes two metrics related to the estimation of uncertainty were further included. One was designed to tackle detection uncertainty and the other segmentation uncertainty. In terms of uncertainty validity, elements are considered as either truly certain (TC), truly uncertain (TU), falsely certain (FC) or falsely uncertain (FU) as per Table 4.

The uncertainty was calculated as $(TU + TC)/(TU + TC + FC + FU)$.

The segmentation uncertainty was only assessed over true positive detected elements, assessing

probabilistic uncertainty accuracy as $\frac{\sum_{TP}(1-Unc)+\sum_{FN+FP}Unc}{TP+FN+FP}$

All metrics were computed per image and the distribution over all cases of the test set was used for the final ranking. For each task, ranking of the methods was performed following the method described for the Medical Image Decathlon challenge (Antonelli et al., 2021). Pairwise comparisons were performed using the Mann-Whitney U-test for the Mean Dice over cases with F1 > 0 and the Wilcoxon paired test for the other metrics due to their non-normal distribution. For each method, the number of times it was found significantly better (with a p-value $\leq$0.05 for significance) than another was used to rank the given metric. The final rank was obtained as the average across the ranks (lower being better). The robustness of the ranking was further assessed using the distribution of Kendall's tau correlation coefficient between ranking for all cases and the one obtained for 1000 bootstrap samples as described in (Wiesenfarth et al., 2021).

To identify the best overall team, the ranks were averaged across all common metrics of all tasks for the teams that provided a solution to all three tasks.

2.6. Additional analyses

Further analyses were performed to inform on the following aspects: 1) clinical performance, 2) performance variability across datasets, 3) regional variability in performance (Task 1 - EPVS), 4) inter rater variability (Task 3 - Lacunes and part of Task 1 - EPVS), and finally ensemble performance using either all methods (EnsembleAll) or the top 50% (EnsembleTop).

Clinical performance. For each task, the most clinically relevant metric was further defined and used to compare the different methods. For Task 1 - EPVS, to emphasize the notion of burden of EPVS, the correlation between predicted and reference volumes across the population of test cases was used. For Task 2 - Microbleeds and Task 3 - Lacunes where a binary statement of existence or absence is most clinically relevant, the balanced accuracy over cases considered as a whole-image classification task was chosen.

Cross-dataset performance. For each task, the performance of each method was additionally computed per dataset and then compared. The ranking was also computed per dataset to examine specific discrepancies between cohorts.

Regional performance. To assess whether the performance of the proposed methods differed depending on the region for Task 1 - EPVS, the evaluation was run for each region (centrum semi-ovale, basal ganglia, hippocampus and mesencephalon) separately. For each method, pairwise comparison across regions was performed to assess whether a given method performed better on a given area. The overall ranking between methods was also computed per region.

Inter-rater variability. For Task 1 - EPVS and Task 3 - Lacunes for which annotations by two raters were available, the evaluation was run considering alternatively each rater as the reference. While the overall absolute differences (volume and number of identified components) between the two raters are independent of the reference chosen (rater 1 or rater 2), changing the reference will affect F1 score and Mean Dice calculation due to differences in definition of true positives.

Ensemble performance. Two ensemble solutions were created and evaluated. The average of all solutions (EnsembleAll) and the average of the predictions from the top 50% in overall rank of the methods (EnsembleTop). EnsembleAll and EnsembleTop were compared to the individual methods for each task. The number of participating teams being 4 for Task 1 - EPVS, EnsembleTop in this case consists in the union of two best performing methods.

3. Results

3.1. Challenge submission and participating teams

Over the period of the challenge, the data set has been requested for 353 downloads. Across the two validation periods, we received requests from 1 team at validation stage 1 and 4 teams at validation stage 2. The final submission of dockerized solutions and their documented description to be applied to the test sets was composed of 4 teams for Task 1 - EPVS, 9 teams for Task 2 - Microbleeds and 6 teams for Task 3 - Lacunes. Only 2 teams participated in all 3 tasks. Table 5 summarizes in which task each team participated.

Table 6 reflects for each task and team the average time needed to evaluate one case, the GPU memory consumption, the docker details for memory requirements (CPU/GPU) and the methods' characteristics. The memory details are presented both as requested by the participants based on their training settings and as measured on a single case allowing for memory flooding. All methods using Stochastic Gradient Descent (SGD) as optimizer applied Nesterov Momentum with value of

Team Name	Task 1 EPVS	Task 2 Microbleeds	Task 3 Lacunes
BigrBrain	✓	✓	✓
Dawai		✓	✓
EMC_N			✓
MixLacune			✓
MixMicrobleed		✓	
MixMicrobleedNet		✓	
Neurophet	✓		✓
TeamTea	✓	✓	✓
Tfff		✓	
TheGPU	✓	✓	
ValdoNN		✓	
Zihao		✓	

Table 5: Participation of the teams across the different tasks

0.99. Poly learning rate scheduling is defined as multiplying the learning rate by $\left(1 - \frac{epoch}{epoch_{max}}\right)^{0.9}$. The following architectures were listed by the participating teams: 2D Unet (Ronneberger et al., 2015), 3D Unet (Çiçek et al., 2016), nnUnet (Isensee et al., 2021), MaskRCNN He et al. (2017), Mask-RetinaNet (Farady et al., 2020), ResNet (He et al., 2016). Beyond the well-known Dice (Milletari et al., 2016) and binary cross-entropy losses, others such as focal loss (Lin et al., 2017) and blob loss (Kofler et al., 2022) were mentioned. Adam (Kingma and Ba, 2014), SGD (Gardner, 1984) and Ranger21 Wright and Demeure (2021) were the optimizers used.

Table 6: Details of the methods of the participating teams for each task. Abbreviations: Aug. - Augmentation; BCE - Binary Cross Entropy; wBCE - weighted Binary Cross-Entropy; CSF - Cerebro spinal fluid; ES - Early Stopping; LR - Learning Rate; MAE - Mean Absolute Error; Mem - Memory; NM - Nesterov Momemtum (value 0.99); Norm. - Normalization; Optim. - Optimizer; PLRS - Poly learning rate schedule; Preproc. - Preprocessing; Pret. - Pretrained; Postproc. - Postprocessing;Req. - Requested; RF - Random Forest; SGD - Stochastic Gradient Descent; Val - Validation

	Team	Time (min)	Mem (GB)	Req	Method	Loss	Dim.	Input	Patches	Preproc.	Optim.	LR	Stopping criterion	PostProc.	Aug.	Val%	Framework	Pret
Task 1 - EPVS	BigrBrain	0.87	1.9	32 10	UNet	Dice	2D	All	225 x 225	Min-max Norm Resampling Cropping	SGD NM	0.01 PLRS	1000 ES 10		Rotation Zooms Shifts Flips	20	Pytorch Ignite	
	Neurophet	1.92	1.7	10 20	MaskRCNN	BCE Focal MAE	2.5D	All		Norm								
	TeamTea	1.38	3.7	10 8	nnUNet	Dice	2D	All	256 x 224	Cropping BF corr. Z-score Norm. Resampling	SGD NM	0.01 PLRS	1000		Zoom Flip Gaussian noise	20		
	TheGPU	8.2	NA	10 0	RF	NA	2D	T2		Cropping min-max Norm.					33			
Task 2 - Microbleeds	BigrBrain	0.9	2.7	32 10	nnUNet	Dice	2D	All	512 x 512	Min-max Norm Resampling Cropping	SGD NM	0.01 PLRS	1000 ES 10		Rotation Zooms Shifts Flips	20	Pytorch Ignite	
	Dawai	11.2	42.2	128 48	UNet	Blob	3D	All	192 x 192 x 32	Quantile Norm.	Ranger 21				Flips Gaussian Noise Affine		MONAI	X
	MixMicrobleed	45.8	43.1	10 10	MaskRCNN UNet	Dice BCE MAE	2D 2.5D	64 x 64 Whole	Z score Norm. Resampling	Adam	0.000005 0.00005	15 50	X	Affine Flips	20		X	
	MixMicrobleedNet	1.4	3.2	10 10	nnUNet	Dice	3D	All			SGD NM	0.01 PLRS				0		
	TeamTea	2	3.3	10 8	UNet	Dice	3D	All	96 x 192 x 128	Z-score Norm BF corr. Resampling Cropping	SGD NM	0.01 PLRS	1000		Zoom Flips Noise	30	nnUNet	
	Tfff	1.5	6.5	10 8	ResNet UNet	Dice	2D	All	320x320	min-max Norm.					Flips Rotation Translation		Pytorch Ignite	
	TheGPU	5	NA	5 0	RF		2D	T2*										
	ValdoNN	0.6	2.3	10 8	nnUNet	Dice BCE	2D	All	512 x 512	Z-score Norm. Resampling Cropping	SGD NM	0.01 PLRS	1000		Rigid Zoom Gaussian noise		nnUNet	
	Zihao	1.6	4.6	8 6	UNet FCN/AlexNet	wBCE Dice	3D	All	20x20x16 24x24x20	z-score Norm Resampling Cropping	SGD NM Adam	0.01 PLRS	150 80 100		Translation Rotation Flips	20% (5)	Pytorch Ignite	
Task 3 - Lacunes	BigrBrain	0.8	2	32 10	UNet	Dice	2D	All	384 x 320	Min-max Norm Resampling Cropping	SGD NM	0.01 PLRS	1000 ES 10		Rotation Zooms Shifts Flips		Pytorch Ignite	
	Dawai	11.9	42.2	128 48	UNet	Blob	3D	All	192 x 192 x 32	Quantile Norm. Cropping	Ranger 21						MONAI	
	EMC_N	32.1	42.7	24 12	UNet	wBCE	3D	All		Resampling BF correction Norm (wrt CSF)	Adam	0.0005	ES(20)	X	Rotation Flips	10%	Keras	X
	MixLacune	9.9	6.1	10 10	MaskRCNN UNet	Dice	2D	All	64 x 64 32 x 32	Z-score Norm.	Adam	0.00005 0.0001	20 30	X	Flips			X
	Neurophet	1.8	1.7	10 20	MaskRetinaNet	BCE Focal MAE	2.5D	All		Norm.	Adam	0.0001	ReduceLR OnPlateau					
	TeamTea	2	3.3	10 8	UNet	Dice	3D	All	96 x 192 x 128	Z-score Norm BF corr. Resampling Cropping	SGD NM	0.01 PLRS	1000		Zoom Flips Noise		nnUNet	

Among all the submissions, only one team (TheGPU) proposed an alternative to a deep learning solution. The majority of the proposed methods were trained as pure segmentation solutions and a few teams submitted a detection+segmentation solution based on Mask-RCNN (He et al., 2017) or Mask Retina net (Farady et al., 2020). Across all tasks, when a deep learning solution was proposed, the UNet architecture was the most common choice. For all three tasks, the time required to process a case and the GPU memory requirements varied greatly. For Task 2 - Microbleeds for instance duration ranged from less than 1 minute to 45.8 min and memory consumption of 2.4 to 43 GB (allowing for memory flooding). In terms of the methodology for uncertainty assessments in Task 3 - Lacunes, the two teams submitting methods to all three tasks did not provide any uncertainty map. Among the 4 remaining teams, most used directly the probabilistic value of their output as measure of uncertainty while mixLacune defined an uncertainty zone at the border of their detected lacunes.

For all teams, key characteristics of the proposed methods are summarized in table 6. Additional details can be found for each team on the OpenReview repository `https://openreview.net/group?id=MICCAI.org/2021/Challenge/VALDO`.

3.2. Metric values

For each task the detection and the segmentation are reported across all teams.

Task 1 - Enlarged Perivascular Spaces (EPVS). The summary statistics for each team and each metric are reported in Table 7.

Table 7: Metrics results for Task 1 - EPVS presented as Median [1st quartile - 3rd quartile] for all metrics. AED - Absolute Element Difference; AVD (in mm3) - Absolute Volume Difference. In bold the significantly best performance across the different teams (excluding the ensemble solutions) and in italic when there is no significant difference compared to the second best.

	Detection		Segmentation	
	F1	AED	Mean Dice	AVD
Bigrbrain	35.81 [28.14 ; 40.42]	*14.50 [6.00 ; 34.50]*	61.09 [55.40 ; 66.57]	45.30 [16.12 ; 89.12]
Neurophet	0.00 [0.00 ; 3.34]	29.00 [13.00 ; 47.00]	28.23 [23.27 ; 29.76]	390.15 [250.72 ; 636.58]
TeamTea	17.12 [6.79 ; 25.90]	41.00 [24.25 ; 69.25]	55.07 [46.25 ; 64.23]	106.05 [73.00 ; 175.86]
TheGPU	**38.92 [28.87 ; 49.44]**	16.00 [9.00 ; 35.75]	**72.38 [64.97 ; 77.12]**	*45.20 [23.79 ; 82.21]*
EnsembleAll	38.62 [28.1 ; 44.82]	24.00 [12.00 ; 46.00]	64.33 [59.14 ; 68.40]	96.15 [63.67 ; 151.69]
EnsembleTop	38.86 [31.19 ; 45.13]	29.00 [15.25 ; 50.25]	67.38 [58.24 ; 72.23]	36.10 [20.15 ; 66.33]

Figure 4 presents the distribution of metrics values for detection (top row) and segmentation metrics (bottom row) for Task 1 - EPVS.

	Detection		Segmentation	
	F1	AED	Mean Dice	AVD
Bigrbrain	16.67 [0.00 ; 36.10]	9.00 [5.00 ; 16.00]	81.17 [71.86 ; 89.69]	52.47 [15.45 ; 171.98]
Dawai	0.00 [0.00 ; 40.00]	1.00 [1.00 ; 3.00]	68.35 [52.99 ; 77.71]	12.40 [6.29 ; 33.05]
MixMicrobleed	0.00 [0.00 ; 0.00]	1e5 [499.5 ; 1e5]	64.36 [55.79 ; 68.58]	1e5 [4728 ; 1e5]
MixMicrobleedNet	**68.42 [36.67 ; 100.00]**	**1.00 [0.00 ; 1.00]**	**84.01 [79.48 ; 87.62]**	**8.77 [2.48 ; 24.30]**
TeamTea	66.67 [0.00 ; 100.00]	**1.00 [0.00 ; 1.00]**	82.57 [74.65 ; 87.50]	11.30 [1.81 ; 25.39]
Tfff	40.00 [18.18 ; 66.67]	3.00 [1.00 ; 6.00]	77.65 [62.43 ; 89.13]	15.27 [4.33 ; 49.33]
TheGPU	0.00 [0.00 ; 0.00]	4.00 [1.00 ; 10.00]	49.46 [36.89 ; 78.14]	602.89 [159 ; 1842.02]
ValdoNN	50.00 [0.00 ; 68.15]	1.00 [1.00 ; 2.00]	80.00 [66.67 ; 87.68]	12.00 [3.14 ; 24.91]
Zihao	66.67 [20.83 ; 100.00]	1.00 [0.00 ; 2.00]	80.00 [73.34 ; 88.04]	9.61 [3.20 ; 21.51]
EnsembleAll	66.67 [0.00 ; 100.00]	1.00 [0.00 ; 1.00]	81.22 [71.35 ; 87.27]	12.87 [4.93 ; 27.26]
EnsembleTop	75.68 [38.18 ; 100.00]	1.00 [0.00 ; 1.00]	77.90 [29.91 ; 87.23]	11.25 [2.81 ; 21.82]

Table 8: Metrics results for Task 2 - Microbleeds presented as Median [1st quartile; 3rd quartile] for each metric. AED - Absolute Element difference; AVD - Absolute volume difference (in mm3). In bold, the significantly best performance per metric across teams (excluding the ensemble solutions)

Task 2 - Microbleeds. Figure 5 presents the distribution of metrics values for detection (top row) and segmentation metrics (bottom row) for Task 2 - Microbleeds with Table 8 presenting the metrics values across all teams.

Task 3 - Lacunes. Table 9 presents the results obtained for Task 3 - Lacunes.

	Detection		**Segmentation**	
	F1	**AED**	**Mean Dice**	**AVD**
BigrBrain	7.69 [5.06 ; 16.49]	27.50 [20.25 ; 33]	40.84 [27.02 ; 50.27]	123.93 [79.49 ; 182.64]
Dawai	15.38 [0.00 ; 25.00]	6.00 [3.00 ; 10.00]	40.09 [26.20 ; 45.31]	78.93 [26.94 ; 209.24]
EMC_N	3.92 [0.00 ; 54.55]	2.00 [1.00 ; 4.75]	20.49 [12.21 ; 34.08]	125.60 [45.08 ; 375.96]
MixLacune	6.25 [0.00 ; 12.00]	22.00 [13.50 ; 26.00]	16.85 [10.31 ; 27.59]	33.95 [16.88 ; 107.69]
Neurophet	4.55 [0.00 ; 10.53]	20.00 [11.50 ; 34.00]	8.82 [3.73 ; 15.33]	471.40 [244.16 ; 891.16]
TeamTea	*28.57 [0.00 ; 57.14]*	**1.00 [0.00 ; 2.00]**	**45.75 [36.74 ; 56.17]**	**14.88 [0.00 ; 40.29]**
EnsembleAll	28.57 [0.00 ; 60.87]	1.00 [0.00 ; 2.00]	37.98 [22.13 ; 44.55]	13.05 [0.07 ; 61.03]
EnsembleTop	30.77 [0.00 ; 66.67]	1.00 [0.00 ; 2.00]	38.17 [25.48 ; 45.26]	9.68 [1.05 ; 63.28]

Table 9: Metrics results for Task 3 - Lacunes presented as median [1st quartile ; 3rd quartile]. AED - Absolute Element difference; AVD - Absolute volume difference (mm3). Bold font indicates best performance across the teams (excluding ensemble solutions) when significantly better than all others. Italic font indicates best performance when not significantly better than the second ranking

while Table 10 shows the metrics for the uncertainty component of the task excluding BigrBrain and TeamTea who did not provide an uncertainty map.

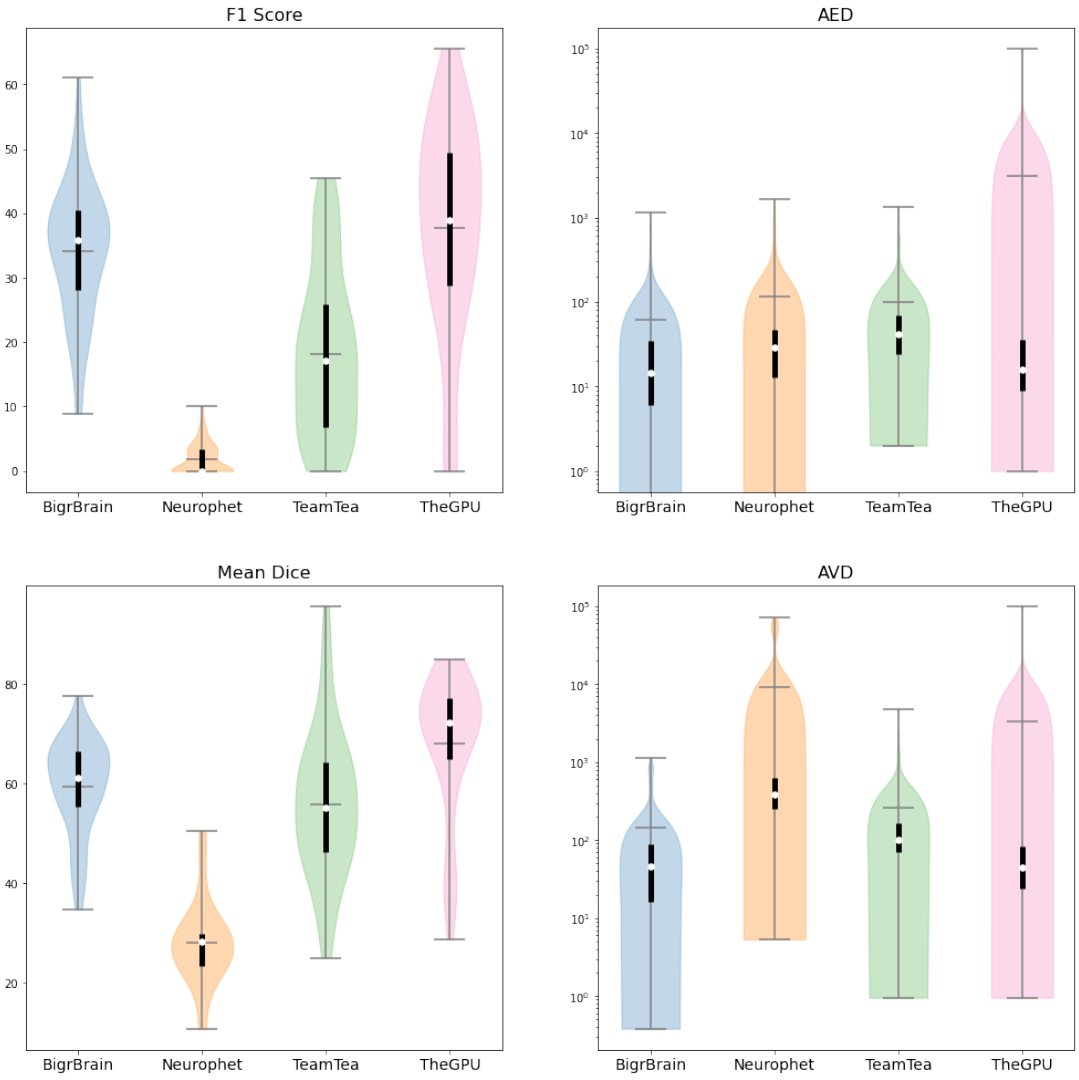

Figure 4: Distribution of metrics values across the different teams for detection metrics (top row) and segmentation metrics (bottom row) for Task 1 - EPVS

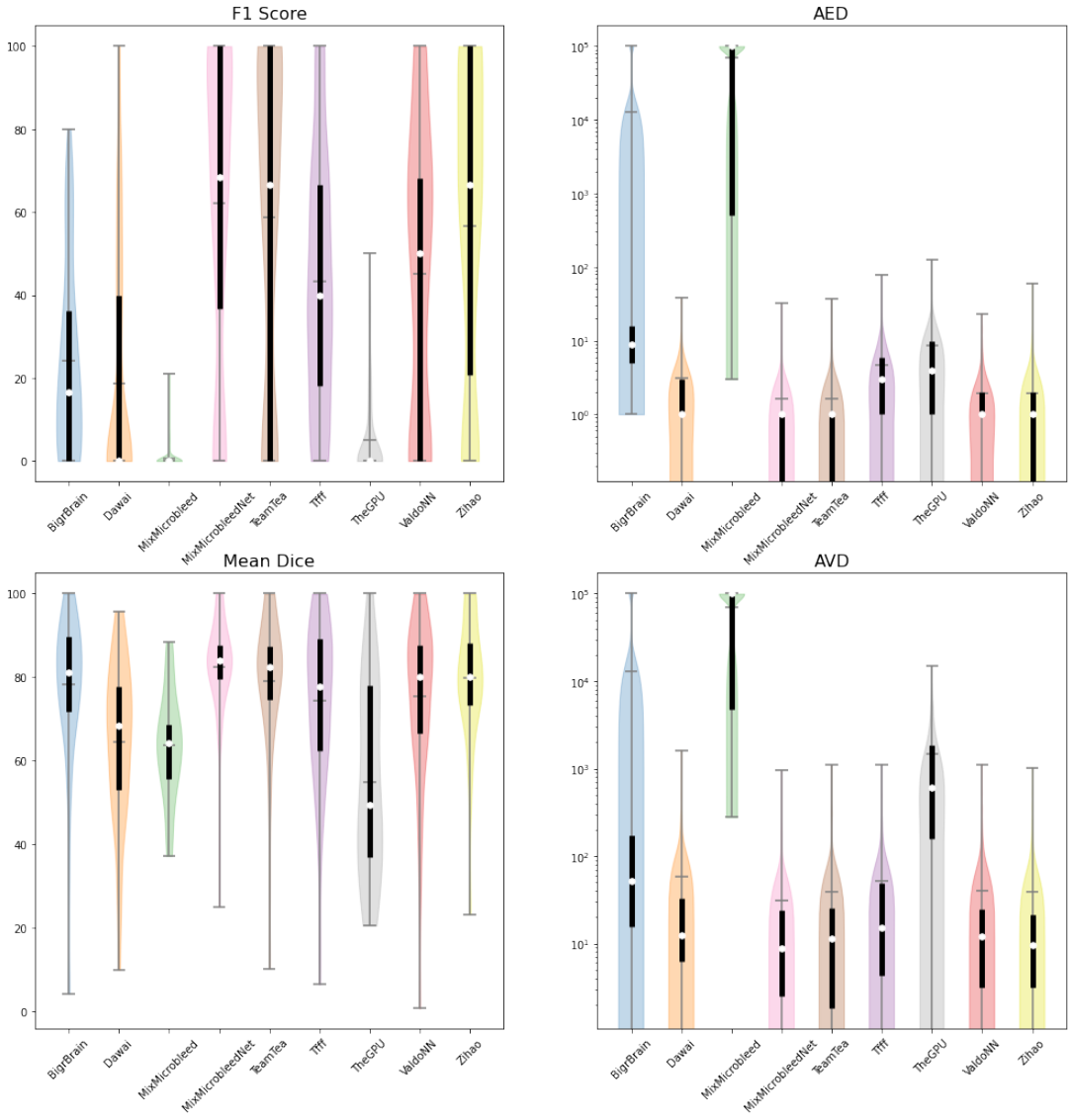

Figure 5: Distribution of metrics values across the different teams for detection metrics (top row) and segmentation metrics (bottom row) for Task 2 - Microbleeds; AED - Absolute Element Difference; AVD - Absolute Volume Difference

Table 10: Metrics related to uncertainty for Task 3 - Lacunes presented as median [1st quartile - 3rd quartile. AED - Absolute Element difference; AVD - Absolute volume difference (in mm3).

	Detection Unc	**Segmentation Unc**
Dawai	0.00 [0.00 ; 25.00]	63.65 [0.00 ; 87.73]
EMC_N	100.00 [86.81 ; 100.00]	0.00 [0.00 ; 67.94]
MixLacune	0.00 [0.00 ; 3.57]	4.76 [0.00 ; 24.39]
Neurophet	0.00 [0.00 ; 6.82]	0.00 [0.00 ; 23.18]

Figure 6 presents the distribution of metrics values for detection (top row) and segmentation metrics (bottom row) for Task 3 - Lacunes.

Figure 7 shows the distribution of metrics values for the assessment of uncertainty applied for Task 3 - Lacunes.

3.3. Rankings

Table 11 presents the overall ranking, according to the number of tasks undergone and for each individual task when relevant.

Table 11: Ranking across all tasks grouped by number of tasks to which each team participated. Across all metrics, D refers to detection and S to segmentation, R to relative, A to absolute and U to uncertainty. DR refers to F1 score, DA to Absolute element difference, SR to Mean Dice, SA to absolute volume difference, DU to detection uncertainty and SU to segmentation uncertainty. Tot is the overall rank for a given task

	Task 1 - EPVS					**Task 2 - Microbleeds**					**Task 3 - Lacunes**						
Team	**DR**	**DA**	**SR**	**SA**	**Tot**	**DR**	**DA**	**SR**	**SA**	**Tot**	**DR**	**DA**	**SR**	**SA**	**DU**	**SU**	**Tot**
TeamTea	3	3.5	3	2.5	3	1.5	2.5	2.5	3	2	1.5	1	1	1			**2**
BigrBrain	2	1.5	2	1	2	6	8	3.5	7	6	4	5.5	2.5	4.5			**6**
Dawai						5	7	7.5	5.5	7	3	3	2.5	3	2	1	**1**
TheGPU	1	1.5	1	2.5	1	7	8	9	8	8							
Neurophet	4	3.5	4	4	4						5.5	5.5	6	6	3	3.5	**5**
MixMicrobleedNet						1.5	1	1	1	1							
Zihao Team						3	2.5	2.5	3	3							
ValdoNN						4	4.5	5	3	4							
Tfff						6	4.5	5	5.5	5							
MixMicrobleed						9	9	7.5	9	9							
EMC_N											1.5	2	4.5	4.5	1	2	**2**
MixLacune											5.5	4	4.5	2	4	3.5	**4**

Table 12 reflects the distribution of Kendall's Tau coefficient when assessing the robustness of

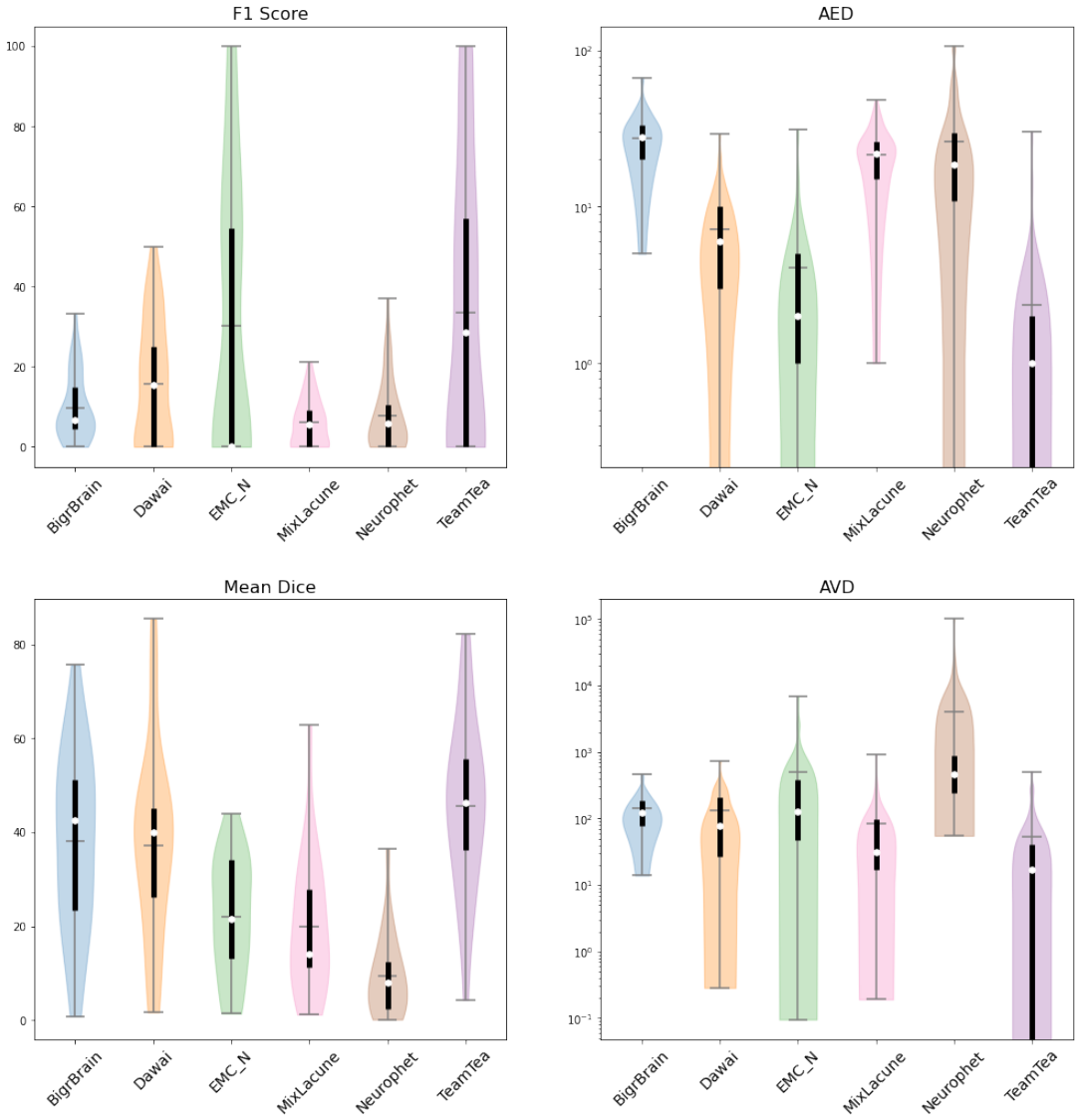

Figure 6: Distribution of metric values across the different teams for detection metrics (top row) and segmentation metrics (bottom row) for Task 3 - Lacunes

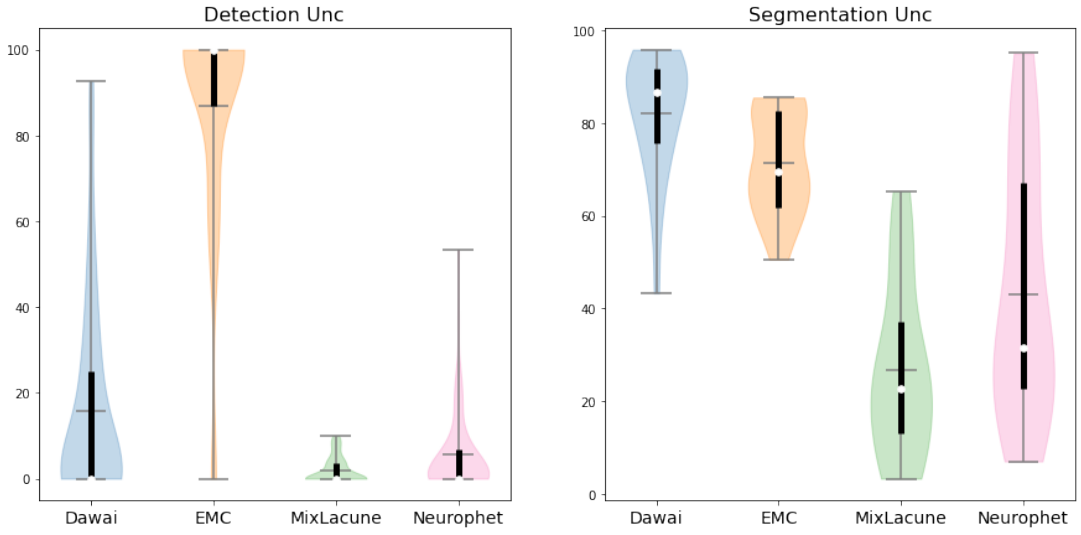

Figure 7: Distribution of metric values across the different teams for the assessment of uncertainty for Task 3 - Lacunes

the ranking for each metric using 1000 bootstrap samples.

Table 12: Distribution characteristics (mean and standard deviation) Kendall's Tau correlation coefficient in % between final ranking and bootstrap samples (1000 samples). Across all metrics, D refers to detection and S to segmentation, R to relative, A to absolute and U to uncertainty. DR refers to F1 score, DA to Absolute element difference, SR to Mean Dice, SA to absolute volume difference, DU to detection uncertainty and SU to segmentation uncertainty.

	DR	DA	SR	SA	DU	SU
Task 1 - EPVS	96.13 (4.33)	93.55 (7.39)	97.87 (4.45)	97.33 (4.02)		
Task 2 - Microbleeds	98.11 (1.81)	98.36 (1.70)	98.19 (2.38)	87.08 (6.62)		
Task 3 - Lacunes	95.88 (6.57)	97.46 (3.98)	94.68 (3.26)	93.19 (5.02)	99.85 (1.13)	95.82 (8.37)

3.4. Additional analyses

3.4.1. Clinical relevant markers

Task 1 - EPVS. For Task 1, since the burden of PVS is currently clinically considered the most valuable insight, the Spearman correlation coefficient between predicted and reference burden across all test cases was calculated for overall volume and element count and is presented in Figure 8 along with the log-transformed relationship between reference and predicted burden in terms of volume (top) and count (bottom).

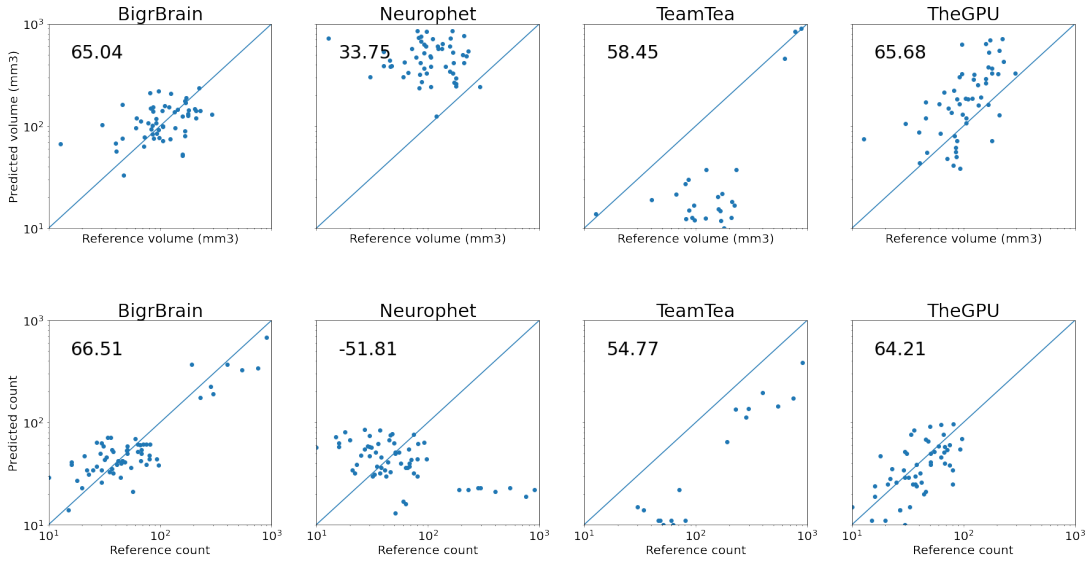

Figure 8: Association between reference and predicted PVS burden across the participating teams for volume (top row) and count (bottom row). The Spearman rho (%) is indicated on each plot.

Task 2 - Microbleeds. For cerebral microbleeds, classifying the absence or presence of any microbleeds was deemed clinically the most relevant assessment. Balanced accuracy over the test set varied from 29.5% for team Dawai to 87.3% for team MixMicrobleed. Figure 9 presents the confusion matrices for each of the teams.

Task 3 - Lacunes. Similarly, Figure 10 shows the confusion matrix for correctly identifying cases that have at least one lacune. For the 6 participating teams, balanced accuracy was close to 0.5 for almost all teams as they predicted the presence of at least one lacune in almost all cases. Only TeamTea was able to recognize cases without lacunes, with 78.3% balanced accuracy.

3.4.2. Cross-dataset variability

Performance varied greatly across datasets, being systematically overall better on RSS dataset than others (SABRE or ALFA). For all three tasks, Figure 11 presents the variation of F1 and Mean Dice score across datasets for all teams and Table 13 presents median and interquartile range for all tasks across datasets for F1 score and Mean Dice.

Ranking varied also slightly across datasets as indicated in Table 14.

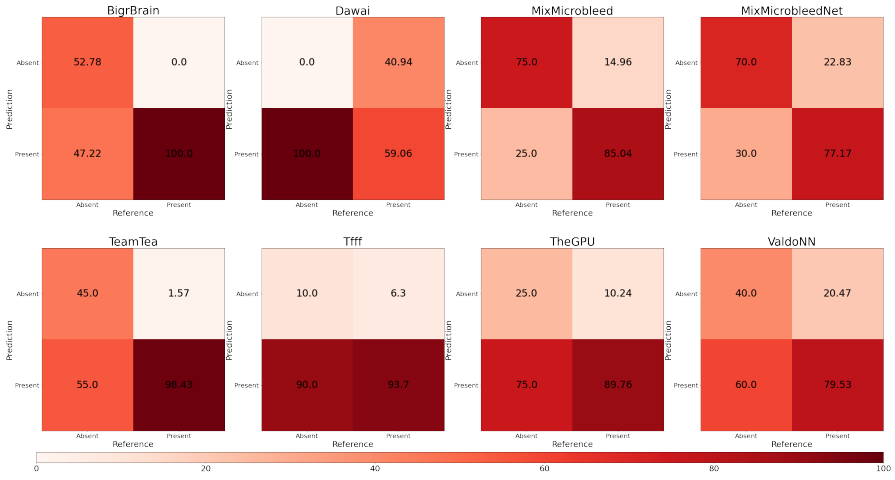

Figure 9: Confusion matrix regarding the classification of an image as containing at least one microbleed based on obtained prediction images.

Table 13: F1 score and Mean Dice presented as median [1st quartile; 3rd quartile] across the different datasets for all three tasks

		F1 Score			Mean Dice		
		ALFA	RSS	SABRE	ALFA	RSS	SABRE
Task 1	BigrBrain		35.40 [27.89 ; 40.36]	38.20 [31.19 ; 40.25]		63.56 [58.05 ; 67.29]	43.10 [41.94 ; 46.92]
	Neurophet		1.95 [0.00 ; 3.71]	0.00 [0.00 ; 0.00]		28.6 [23.90 ; 30.17]	NA
	TeamTtea		13.50 [6.02 ; 21.76]	34.80 [25.10 ; 40.93]		59.66 [50.08 ; 65.55]	45.05 [42.88 ; 46.91]
	TheGPU		43.71 [34.55 ; 51.67]	24.32 [2.61 ; 28.42]		73.41 [69.68 ; 78.6]	34.96 [7.16 ; 39.25]
Task 2	BigrBrain	11.11 [0.00 ; 16.67]	30.77 [13.81 ; 51.47]	36.36 [21.81 ; 58.24]	82.21 [73.3 ; 93.32]	75.07 [69.41 ; 81.84]	90.13 [84.63 ; 92.19]
	Dawai	0.00 [0.00 ; 0.00]	41.43 [25.00 ; 66.67]	0.00 [0.00 ; 0.00]	57.20 [43.49 ; 70.91]	69.47 [53.16 ; 77.82]	63.33 [56.31 ; 69.23]
	Mixmicrobleed	0.00 [0.00 ; 0.00]	0.00 [0.00 ; 0.75]	0.00 [0.00 ; 0.42]	0.00 [0.00 ; 0.00]	58.90 [54.56 ; 67.04]	68.62 [66.67 ; 79.41]
	MixmicrobleedNet	66.67 [0.00 ; 100]	77.81 [66.67 ; 100.00]	51.67 [50.00 ; 69.23]	87.18 [74.71 ; 96.67]	82.79 [79.82 ; 85.20]	84.21 [79.35 ; 87.39]
	TeamTea	50.00 [0.00 ; 100.00]	80.00 [66.67 ; 100.00]	50.00 [30.22 ; 68.75]	85.16 [65.38 ; 100.00]	82.08 [77.83 ; 85.24]	84.62 [74.65 ; 87.66]
	Tfff	20.00 [7.68 ; 40.00]	65.15 [47.50 ; 76.41]	40.00 [33.33 ; 55.91]	80.00 [63.19 ; 99.46]	68.58 [57.64 ; 80.72]	86.19 [79.14 ; 89.46]
	TheGPU	0.00 [0.00 ; 0.00]	0.00 [0.00 ; 9.95]	0.00 [0.00 ; 10.01]	79.43 [56.53 ; 83.76]	40.00 [32.63 ; 48.54]	66.67 [53.28 ; 79.70]
	ValdoNN	0.00 [0.00 ; 66.67]	66.67 [38.82 ; 80.00]	50.00 [32.14 ; 51.56]	86.06 [70.24 ; 100.00]	70.91 [62.08 ; 81.48]	87.00 [81.51 ; 89.58]
	Zihao	50.00 [0.00 ; 100.00]	74.81 [66.67 ; 94.23]	45.00 [22.92 ; 63.54]	87.71 [80.00 ; 100.00]	76.98 [71.69 ; 80.46]	85.42 [74.53 ; 88.85]
Task 3	BigrBrain		7.41 [5.48 ; 14.91]	8.39 [3.80 ; 17.75]		42.81 [27.39 ; 51.80]	30.17 [22.30 ; 37.37]
	Dawai		20.00 [0.00 ; 33.33]	0.00 [0.00 ; 0.00]		39.88 [25.38 ; 44.82]	57.14 [57.14 ; 57.14]
	EMC_N		44.44 [0.00 ; 66.67]	0.00 [0.00 ; 0.00]		21.42 [13.09 ; 34.24]	2.20 [2.20 ; 2.20]
	MixLacune		6.25 [0.00 ; 10.81]	9.09 [0.00 ; 16.16]		16.85 [10.31 ; 28.04]	15.45 [12.51 ; 19.3]
	Neurophet		6.25 [0.00 ; 14.29]	0.00 [0.00 ; 0.46]		8.82 [4.98 ; 14.65]	11.37 [6.86 ; 15.88]
	TeamTea		40.00 [0.00 ; 66.67]	0.00 [0.00 ; 1.61]		45.75 [34.58 ; 55.75]	64.66 [54.40 ; 74.92]

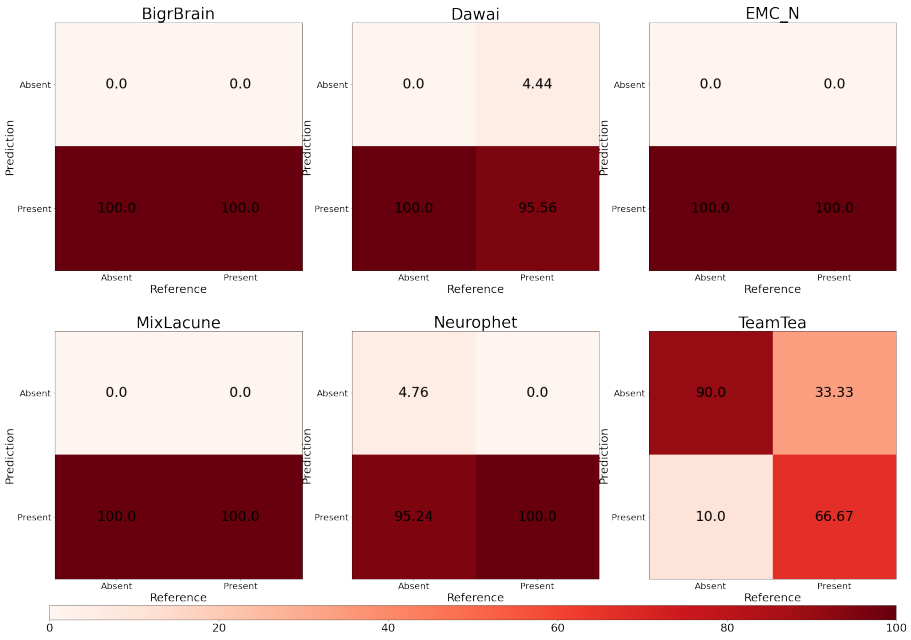

Figure 10: Confusion matrix regarding the classification of an image as containing at least one lacune based on obtained prediction images.

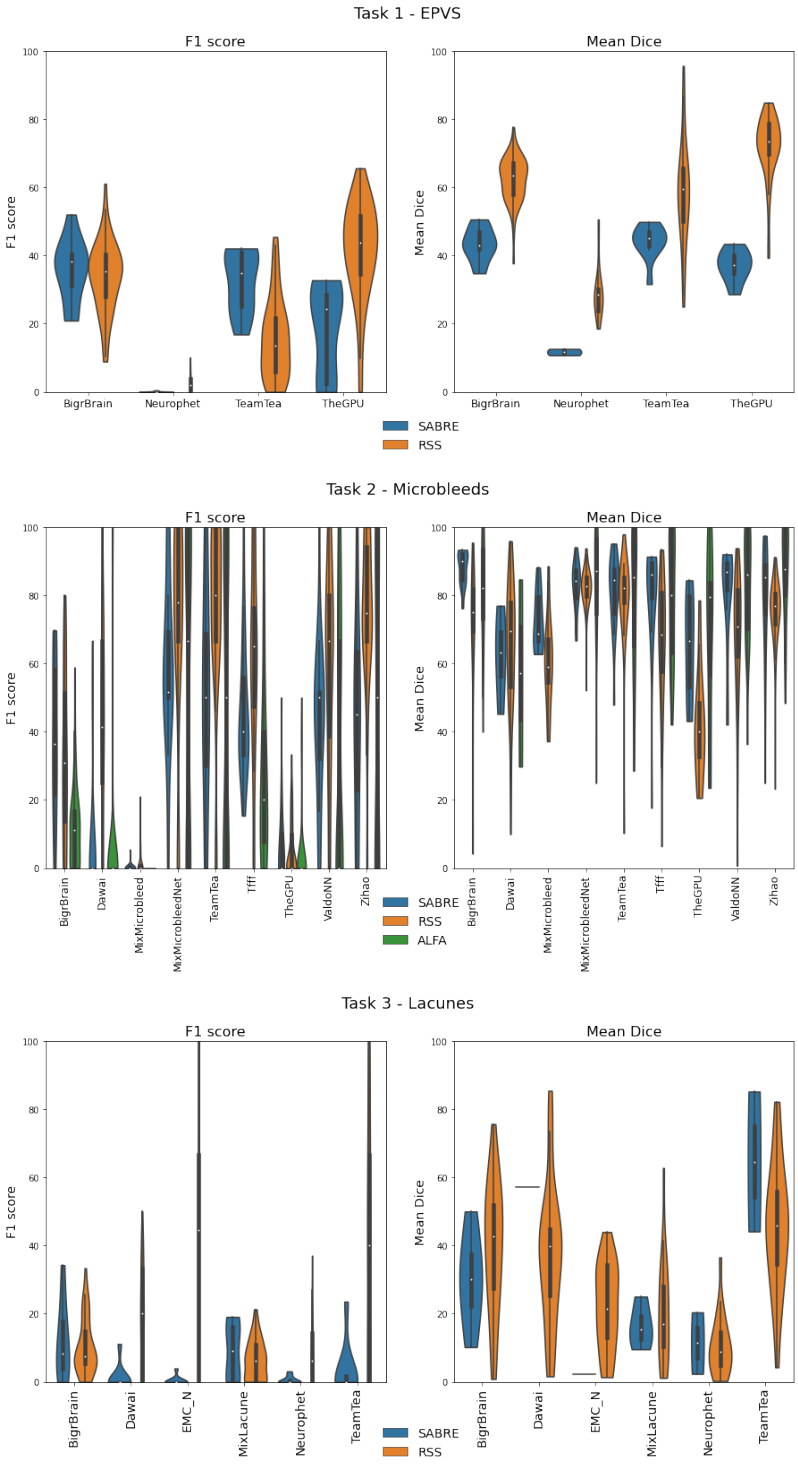

Figure 11: Distribution of results for F1 (left column) and Mean Dice (right column) across different datasets for the three tasks (each row represents one task).

Table 14: Ranking calculated for each dataset separately

		ALFA	RSS	SABRE
Task 1 - EPVS	**BigrBrain**		2	1
	Neurophet		4	4
	TeamTea		3	2
	TheGPU		1	3
Task 2 - Microbleeds	**BigrBrain**	7	7	6
	Dawai	6	6	7
	MixMicrobleed	9	9	9
	MixMicrobleedNet	2	1	2
	TeamTea	3	2	1
	Tfff	5	5	5
	TheGPU	8	8	8
	ValdoNN	4	4	3
	Zihao	1	3	4
Task 3 - Lacunes	**BigrBrain**		5	5
	Dawai		2.5	1
	EMC_N		1	3
	Mixlacune		4	2
	Neurophet		6	6
	TeamTea		2.5	4

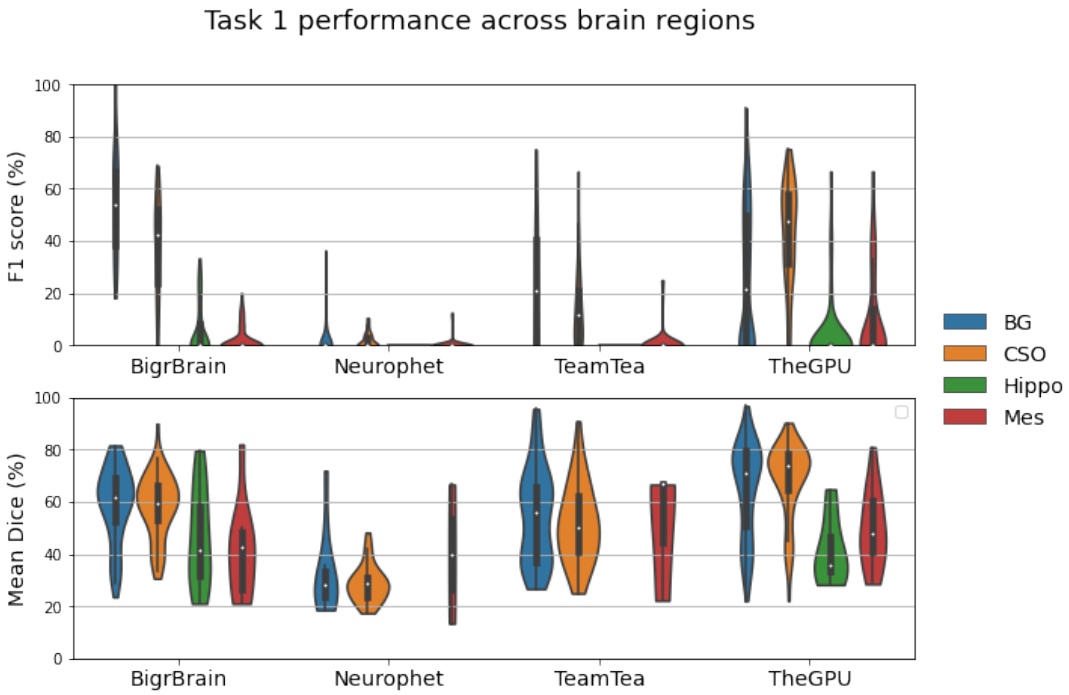

Figure 12: F1 and Mean Dice distribution across the different brain regions for Task 1 - EPVS

3.4.3. Regional variability

Metrics variability for Task 1 - EPVS across different brain regions is illustrated for F1 and Mean Dice in Figure 12.

3.4.4. Inter-rater variability

Inter-rater variability was investigated for tasks and datasets for which two raters provided annotation for the same case (Task 1 - EPVS SABRE dataset, Task 3 - Lacunes all datasets) and results are presented in Table 15.

For Task 1 - EPVS, intra-rater detection was slightly lower than the best method but the inter-rater segmentation performance appeared to be better by quite a strong margin reaching 59.49% in comparison to the best method at 45.5%. The detection performance was notably higher for Task 3 - Lacunes with segmentation performance on par with the best performing method.

Table 15: Metrics values (median [1st quartile - 3rd quartile] presented for the cases where a double rating was available in the test set.

	Detection			Segmentation		
	F1 score R1	**F1 score R2**	**AED**	**Mean Dice R1**	**Mean Dice R2**	**AVD**
Task 1 - EPVS	19.57 [13.48 ; 23.81]	19.86 [13.58 ; 23.81]	135.00 [96.25 ; 316.50]	52.63 [52.07 ; 54.51]	54.07 [51.43 ; 55.05]	651.00 [371.75 ; 2819.75]
Task 3 - Lacunes	48.45 [39.01 ; 61.88]	55.84 [0.00 ; 86.36]	0.50 [0.00 ; 1.00]	59.03 [43.72 ; 64.47]	59.49 [44.95 ; 65.88]	21.51 [0.00 ; 43.63]

3.4.5. Ensembles

For the creation of EnsembleTop, Task 1 - EPVS used predictions from team TheGPU and BigrBrain, Task 2 - Microbleeds used predictions from MixMicrobleedNet, TeamTea, Zihao, and ValdoNN, while for Task 3 - Lacunes, predictions from Dawai, TeamTea and EMC_N were used. Table 16 presents the values of the metrics and the corresponding ranking obtained for each type of ensemble (EnsembleAll, the average of all solutions, and EnsembleTop, the average of the top 50%) across the three tasks. When considering the clinical metrics, performance was higher for both ensemble solutions in Task 1 - EPVS reaching a correlation coefficient of 70.0% and 74.8% for EnsembleAll and EnsembleTop respectively for the count and 69.5 and 80.0% for the volume. For Task 2 - Microbleeds, balanced accuracy was of 77.0% for EnsembleAll and 79.6% for EnsembleTop ranking fourth and third compared to all the teams. Finally, for Task 3 - Lacunes, balanced accuracy reached 75.0% for EnsembleAll, down to 65.3$ for EnsembleTop slightly lower than the 78.0% obtained by TeamTea.

Table 16: Metrics value presented as median [IQR] for the 4 common metrics across the different ensemble types for the three tasks along with associated ranking

		F1	**AD**	**Mean Dice**	**AVD**
Task 1 - EPVS	EnsembleAll	38.62 [28.10 ; 44.82] 3.5	24.00 [12.00 ; 46.00] 3.5	64.33 [59.14 ; 68.40] 4	96.15 [63.67 ; 151.69] 2
	EnsembleTop	38.86 [31.19 ; 45.13] 1.5	29.00 [15.25 ; 50.25] 3.5	67.38 [58.24 ; 72.23] 2	36.10 [20.15 ; 66.33] 2
Task 2 - Microbleeds	EnsembleAll	66.67 [0.00 ; 100.00] 4	1.00 [0.00 ; 1.00] 3	81.22 [71.35 ; 87.27] 6.5	12.87 [4.93 ; 27.26] 7
	EnsembleTop	75.68 [38.18 ; 100] 1	1.00 [0.00 ; 1.00] 1	77.90 [29.91 ; 87.23] 3	11.25 [2.81 ; 21.82] 3
Task 3 - Lacunes	EnsembleAll	28.57 [0.00 ; 60.87] 2.5	1.00 [0.00 ; 2.00] 2	37.98 [22.13 ; 44.55] 3.5	13.05 [0.07 ; 61.03] 2
	EnsembleTop	30.77 [0.00 ; 66.67] 2.5	1.00 [0.00 ; 2.00] 2	38.17 [25.48 ; 45.26] 3.5	9.68 [1.05 ; 63.28] 2

4. Discussion

This manuscript reports the design and outcome of the *"Where is VALDO?"* challenge that took place as a satellite event of MICCAI 2021. Detection and segmentation of three types of markers of cerebral small vessel disease were evaluated as three distinct tasks namely enlarged perivascular spaces (Task 1), cerebral microbleeds (Task 2) and lacunes (Task 3). Among the 12 distinct participating teams, 9 teams provided a solution for Task 2 and 2 teams competed across all three tasks.

Although the challenge was designed to address both detection and segmentation aspects, most of the proposed solutions were designed with a segmentation purpose only - the detection performance considered as a by-product of the prediction. This choice may have been influenced partially by the guidelines to provide only the probabilistic segmentation map that was then post-processed to identify the individual connected components instead of requesting instance segmentation and predicted detections as outputs. However, this strategy appeared to generally work well with segmentation performance being on par with detection performance across all three tasks. Interestingly, there was no strong relationship between memory, time expenditure and overall performance with some of the most greedy methods having lower performance than some of the most cost-effective solutions.

Across all tasks, one team proposed a solution not relying on deep-learning and their strategy had the best performance for Task 1 - EPVS possibly because of the fact that EPVS may be relatively easy to characterise in terms of signal and shape signature. However, none of the proposed methods for Task 1 - EPVS made use of the weak annotation data (count on slices). Also, while some methods only used annotated slices, performance may have been lowered by the absence of use of the masks when only specific parts of a given axial slice were annotated (RSS Data). Most deep learning solutions described using a UNet style architecture at one point of their pipeline either as main network for one-stage methods or for the segmentation component for multi-stage solutions. Interestingly, despite four teams describing the use of the nnUNet (Isensee et al., 2021) architecture for Task 2 - Microbleeds, performance varied greatly across these teams with rank 1, 2, 5 and 7 out of 9. This could potentially be explained by the choice of input data, the dimensionality, or the framework chosen. In the context of microbleeds, using 3D information may be particularly relevant to avoid mimics. This observation highlights the importance of all these steps in the design of a relevant solution, the use of the whole extent of the training data being a key component of the

winner's method. Such consideration is particularly relevant when dealing with a modest number of training examples. When considering choices of augmentation, those involving local changes to input images and/or reference annotation (interpolation, intensity changes, spatial deformation) may cause inconsistencies in the case of very small objects of interest.

In terms of dataset origins, performance was generally higher for the dataset with the highest resolution which was also for Task 2 - Microbleeds and Task 3 - Lacunes the dataset with the highest number of training cases. This is naturally expected as a direct impact on resolution on evaluation metrics and as an overfitting related property.

The amount of training data (in terms of examples of lesions) appeared also to be relevant when comparing the performance of the methods of Task 1 across the different regions of interest, the regions with the most EPVS (centrum semi-ovale and basal ganglia) being the ones with the highest performance across all methods. This may not only be due to the sheer amount of training data in the remaining regions (hippocampus and mesencephalon) but also to the characteristics of the imaging sequences in these regions and the likelihood for mimics (cysts) and higher variability in presentation. Knowledge of the differences in performance across regions is particularly interesting clinically when associations with risk factors and or clinical function have been made specifically in specific anatomical regions in relation to Alzheimer's Disease (Jiménez-Balado et al., 2018) and Parkinson's disease (Duker and Espay, 2007). For Task 1 - EPVS, even for the best teams, the performance presented a large variability which would make their adoption in clinical practice difficult. The overall good correlation between expected and predicted burden may however already be enough to make these tools valuable when investigating associations at population level. For Task 2 - Microbleeds, it appeared that, when correctly detected, the segmentation of lesions was very good. However, even in the best of teams there were issues at the detection level with both cases missed and cases wrongly considered as containing at least one microbleed. The best teams indicated very few lesions which would be relatively practical to visually inspect and reject if necessary. It is here the absence of a systematic bias towards overcall or undercall could make it difficult to integrate in clinical pipelines. For Task 3 - Lacunes, performance appeared quite poor on both detection and segmentation metrics, with a general large overcall of lacunes and when detecting them correctly a lower segmentation performance than for Task 2 - Microbleeds. Such performance would require too much time for editing and checking to be adopted in both clinical practice and research studies.

When comparing the performance across all three tasks, it appeared that the performance was higher on tasks for which the variability in element appearance was lower (EPVS with linear shapes and microbleeds with spherical shapes compared to lacunes with more heterogeneous shapes). The metrics investigated as closest to the current clinical measures of interest were generally in agreement with the overall ranking of the challenge but showed stark differences in terms of clinical viability of the proposed solutions. While for Task 1 - EPVS and Task 2 - Microbleeds the proposed solutions achieved reasonable performance in terms of "clinical" metric, only one team performed reasonably well for Task 3 - Lacunes, with all other solutions systematically finding many lacunes even when there were none. This may be due to the large variability in appearance (i.e. shape, location, intensity signature) as well as the lower number of examples of this type of lesions when compared to those of Task 1 - EPVS and Task 2 - Microbleeds. With all solutions generally producing many false positives, the time required to go through each case and reject many wrongly detected lesion candidates would be prohibitive for clinical adoption. One must however keep in mind that none of these solutions were optimized for this metric and may have performed differently otherwise. In this case the addition of auxiliary tasks in the learning framework to abide to a priori knowledge of burden distribution or to directly optimize such metrics may have interesting results.

In a field where adequate research biomarkers have yet to be properly defined and proven to be reliable (Smith et al., 2019), these observations regarding clinical metrics may lead to define different tasks and solutions for the targeted markers according to their purpose: clinical practice or research. While location, individual volume and shape information may become of interest in the research context as potential new biomarkers, thereby highlighting segmentation as an interesting end-goal, these characteristics may not be yet relevant in the clinical context. In clinical practice, one could imagine a two-stage pipeline with 1) whole-image level classification favouring sensitivity for the flagging of scans where an assessment is required for the presence or absence of a specific marker 2) Specification of lesion location (if needed) for the scans that have been flagged as containing a marker. This second step may be particularly relevant when supporting diagnosis (e.g., distinction between amyloid angiopathy and hypertensive pathology according to microbleed location) or to the explanation of the clinical presentation (e.g., lacune on crucial white matter tract).

A key aspect, not measured here, is the ability of the proposed methods to be used in clinical settings with scans likely to be of lower resolution and to have more artefacts as well as present simultaneously other markers of pathology (e.g stroke, tumours). With the continuous progress in

acquisition protocols and the democratization of scanning abilities, research-grade scanning protocols such as those used in this challenge may become available routinely, thereby limiting issues of protocol related generalizability. However, cohort-related bias may be more difficult to overcome. In fact, in the challenge, data came only from population cohorts and did not include patients with dementia as would be frequent in memory clinics. While efforts were made to provide training examples from the whole spectrum of lesion burden, specific pathological presentations may be missing and the generalizability of the proposed solutions would need to be assessed in these contexts.

Conclusion. In this challenge assessing the current segmentation and detection performance of three markers of cerebral small vessel disease, namely EPVS, Microbleeds and Lacunes, methods targeting directly the segmentation were often quite successful in detecting these small structures. Number of elements on which to train the solutions was strongly predictive of performance, both across tasks and regionally. Manually engineered features became in the case of EPVS relevant enough to compete with deep-learning based strategies. Strikingly, all the presented methods proposed a training based on dense labelling, discarding the weak labelling available for Task 1 - EPVS. While for Task 1 - EPVS and Task 2 - Microbleeds some demonstrated they could potentially be used for population-based research, the large variability in performance across cases may require lengthy visual censoring if they were to be used for individual cases. In this context, it could be relevant to further include the evaluation of performance variability in the assessed tasks. In addition, systematic assessment of prediction confidence (as proposed with the uncertainty metrics of Task 3 - Lacunes) would be of interest for the design of practical implementation.

Funding. The challenge prizes were provided by Nvidia and Icometrix. The SABRE study was funded at baseline by the Medical Research Council, Diabetes UK, and British Heart Foundation and at follow-up by the Wellcome Trust (082464/Z/07/Z), British Heart Foundation (SP/07/001/23603, PG/08/103, PG/12/29/29497 and CS/13/1/30327) and Diabetes UK(13/0004774). The Rotterdam Scan Study is supported by the Erasmus MC University Medical Center, the Erasmus University Rotterdam, the Netherlands Organization for Scientific Research (NWO) Grant 918-46-615, the Netherlands Organization for Health Research and Development (ZonMW), the Research Institute for Disease in the Elderly (RIDE), and the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement No. 601055, VPH-DARE@IT, the Dutch Technology

Foundation STW (Perspectief programme: Population Imaging Genetics The ALFA study is supported by the La Caixa Foundation. CHS is funded by an Alzheimer's Society Junior Fellowship (AS-JF-17-011). KVW and SC are supported by the Deep Learning for Medical Image Analysis (DLMedIA) (project no. P15-26), funded by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO) and which is partly funded by the Ministry of Economic Affairs, with co-financing by Quantib. FD was funded by Netherlands Organisation for Health Research and Development 104003005. BM, BW and FK are supported through the SFB 824, subproject B12, supported by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. IE is funded by DComEX (Grant agreement ID: 956201). BM acknowledges support by the Helmut Horten Foundation. MdG is an employee of, and holds shares in GSK. GSK had no role in the design of this challenge, or the interpretation of the results. MdB is supported by Netherlands Organisation for Scientific Research (NWO) project VI.C.182.042. SI and LL have received funding from the Innovative Medicines Initiative 2 Joint Undertaking under Amyloid Imaging to Prevent Alzheimer's Disease (AMYPAD) grant agreement No. 115952 and European Prevention of Alzheimer's Dementia (EPAD) grant No. 115736. This Joint Undertaking receives the support from the European Union's Horizon 2020 Research and Innovation Programme and EFPIA. HJK was supported by the Galen and Hilary Weston Foundation under the Novel Biomarkers 2019 scheme (№UB190097). JLM is currently a full-time employee of Lundbeck and has served previously as a consultant or on advisory boards for the following for-profit companies, or has given lectures in symposia sponsored by the following for profit companies: Roche Diagnostics, Genentech, Novartis, Lundbeck, Oryzon, Biogen, Lilly, Janssen, Green Valley, MSD, Eisai, Alector, BioCross, GE Healthcare, and ProMIS Neurosciences. JDG is supported by the Spanish Ministry of Science and Innovation (RYC-2013-13054), has received research support from GE Healthcare, Roche Diagnostics and Hoffmann-La Roche and speaker's fees from Biogen and Philips.

Acknowledgements. We are particularly thankful to all participants of the ALFA, RSS and SABRE study. We also would like to thank the team of GrandChallenge.org for their technical support and guidance. The ALFA group study is composed of Müge Akinci, Eider M Arenaza-Urquijo, Annabella Beteta, Anna Brugulat-Serrat, Raffaele Cacciaglia, Alba Cañas, Irene Cumplido, Carme Deulofeu, Ruth Dominguez, Maria Emilio, Carles Falcon, Karine Fauria, Sherezade Fuentes, José

Maria González de Echavarri-Gómez, Oriol Grau-Rivera, Laura Hernandez, Gema Huesa, Jordi Huguet, Paula Marne, Marta Milà-Alomà, Tania Menchón, Carolina Minguillon, Arcadi Navarro, Grégory Operto, Eva M Palacios, Eleni Palpatzis, Cleofé Peña-Gómez, Albina Polo, Sandra Pradas, Blanca Rodríguez-Fernández, Aleix Sala-Vila, Gonzalo Sánchez-Benavides, Gemma Salvadó, Mahnaz Shekari, Anna Soteras, Laura Stankeviciute, Marc Suárez-Calvet, Marc Vilanova, Natalia Vilor-Tejedor.

References

Adams, H.H., Hilal, S., Schwingenschuh, P., Wittfeld, K., van der Lee, S.J., DeCarli, C., Vernooij, M.W., Katschnig-Winter, P., Habes, M., Chen, C., et al., 2015. A priori collaboration in population imaging: the uniform neuro-imaging of virchow-robin spaces enlargement consortium. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1, 513–520.

Alistair, D.G., 2002. Hypertensive cerebral small vessel disease and stroke. Brain pathology 12, 358–370.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., AnnetteKopp-Schneider, Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Huisman, H., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbelaez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, N., Kim, I., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2021. The medical segmentation decathlon. `arXiv:2106.05735`.

Atlason, H.E., Love, A., Sigurdsson, S., Gudnason, V., Ellingsen, L.M., 2019. Segae: Unsupervised white matter lesion segmentation from brain mris using a cnn autoencoder. NeuroImage: Clinical 24, 102085.

Buch, S., Cheng, Y.C.N., Hu, J., Liu, S., Beaver, J., Rajagovindan, R., Haacke, E.M., 2017. Determination of detection sensitivity for cerebral microbleeds using susceptibility-weighted imaging. NMR in biomedicine 30, e3551.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Cordonnier, C., Potter, G.M., Jackson, C.A., Doubal, F., Keir, S., Sudlow, C.L., Wardlaw, J.M., Salman, R.A.S., 2009. Improving interrater agreement about brain microbleeds: development of the brain observer microbleed scale (bombs). Stroke 40, 94–99.

De Boer, R., Vrooman, H.A., Van Der Lijn, F., Vernooij, M.W., Ikram, M.A., Van Der Lugt, A., Breteler, M.M., Niessen, W.J., 2009. White matter lesion extension to automatic brain tissue segmentation on mri. Neuroimage 45, 1151–1161.

Duker, A.P., Espay, A.J., 2007. Parkinsonism associated with striatal perivascular space dilation. Neurology 68, 1540–1540.

Farady, I., Lin, C.Y., Rojanasarit, A., Prompol, K., Akhyar, F., 2020. Mask classification and head temperature detection combined with deep learning networks, in: 2020 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP), IEEE. pp. 74–78.

Gardner, W.A., 1984. Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique. Signal processing 6, 113–133.

Giau, V.V., Bagyinszky, E., Youn, Y.C., An, S.S.A., Kim, S.Y., 2019. Genetic factors of cerebral small vessel disease and their potential clinical outcome. International journal of molecular sciences 20, 4298.

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M.d.C., Dickie, D.A., Wardlaw, J., et al., 2018. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. NeuroImage: Clinical 17, 918–934.

Haffner, C., Malik, R., Dichgans, M., 2016. Genetic factors in cerebral small vessel disease and their impact on stroke and dementia. Journal of Cerebral Blood Flow & Metabolism 36, 158–171.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

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

Ikram, M.A., Brusselle, G., Ghanbari, M., Goedegebure, A., Ikram, M.K., Kavousi, M., Kieboom, B.C., Klaver, C.C., de Knegt, R.J., Luik, A.I., et al., 2020. Objectives, design and main findings until 2020 from the rotterdam study. European journal of epidemiology 35, 483–517.

Ikram, M.A., van der Lugt, A., Niessen, W.J., Koudstaal, P.J., Krestin, G.P., Hofman, A., Bos, D., Vernooij, M.W., 2015. The rotterdam scan study: design update 2016 and main findings. European journal of epidemiology 30, 1299–1315.

Ingala, S., Mazzai, L., Sudre, C.H., Salvadó, G., Brugulat-Serrat, A., Wottschel, V., Falcon, C., Operto, G., Tijms, B., Gispert, J.D., et al., 2020. The relation between apoe genotype and cerebral microbleeds in cognitively unimpaired middle-and old-aged individuals. Neurobiology of Aging 95, 104–114.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., et al., 2019. Automated brain extraction of multisequence mri using artificial neural networks. Human brain mapping 40, 4952–4964.

Jiménez-Balado, J., Riba-Llena, I., Garde, E., Valor, M., Gutiérrez, B., Pujadas, F., Delgado, P., 2018. Prevalence of hippocampal enlarged perivascular spaces in a sample of patients with hypertension and their relation with vascular risk factors and cognitive function. Journal of Neurology, Neurosurgery & Psychiatry 89, 651–656.

Kester, M.I., Goos, J.D., Teunissen, C.E., Benedictus, M.R., Bouwman, F.H., Wattjes, M.P., Barkhof, F., Scheltens, P., van der Flier, W.M., 2014. Associations between cerebral small-vessel disease and alzheimer disease pathology as measured by cerebrospinal fluid biomarkers. JAMA neurology 71, 855–862.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kofler, F., Shit, S., Ezhov, I., Fidon, L., Horvath, I., Al-Maskari, R., Li, H., Bhatia, H., Loehr, T., Piraud, M., Erturk, A., Kirschke, J., Peeken, J., Vercauteren, T., Zimmer, C., Wiestler, B., Menze, B., 2022. blob loss: instance imbalance aware loss functions for semantic segmentation. URL: https://arxiv.org/abs/2205.08209, doi:10.48550/ARXIV.2205.08209.

Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE transactions on medical imaging 38, 2556–2568.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. Medical image analysis 35, 250–269.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., et al., 2020. Bias: Transparent reporting of biomedical image analysis challenges. Medical image analysis 66, 101796.

Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W.H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A., et al., 2015. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. Computational intelligence and neuroscience 2015.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging 34, 1993–2024.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), IEEE. pp. 565–571.

Modat, M., Cash, D.M., Daga, P., Winston, G.P., Duncan, J.S., Ourselin, S., 2014. Global image registration using a symmetric block-matching approach. Journal of medical imaging 1, 024003.

Molinuevo, J.L., Gramunt, N., Gispert, J.D., Fauria, K., Esteller, M., Minguillon, C., Sánchez-Benavides, G., Huesa, G., Morán, S., Dal-Ré, R., et al., 2016. The alfa project: a research platform to identify early pathophysiological features of alzheimer's disease. Alzheimer's & Dementia: Translational Research & Clinical Interventions 2, 82–92.

Østergaard, L., Engedal, T.S., Moreton, F., Hansen, M.B., Wardlaw, J.M., Dalkara, T., Markus, H.S., Muir, K.W., 2016. Cerebral small vessel disease: capillary pathways to stroke and cognitive decline. Journal of Cerebral Blood Flow & Metabolism 36, 302–325.

Potter, G.M., 2011. Neuroimaging of cerebral small vessel disease .

Rensma, S.P., van Sloten, T.T., Launer, L.J., Stehouwer, C.D., 2018. Cerebral small vessel disease and risk of incident stroke, dementia and depression, and all-cause mortality: a systematic review and meta-analysis. Neuroscience & Biobehavioral Reviews 90, 164–173.

Ritter, F., Boskamp, T., Homeyer, A., Laue, H., Schwier, M., Link, F., Peitgen, H.O., 2011. Medical image analysis. IEEE pulse 2, 60–70.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Sigurdsson, S., Aspelund, T., Kjartansson, O., Gudmundsson, E., Jonsson, P.V., van Buchem, M.A., Gudnason, V., Launer, L.J., 2022. Cerebrovascular risk-factors of prevalent and incident brain infarcts in the general population: the ages-reykjavik study. Stroke 53, 1199–1206.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE transactions on medical imaging 17, 87–97.

Smith, E.E., Biessels, G.J., De Guio, F., De Leeuw, F.E., Duchesne, S., Düring, M., Frayne, R., Ikram, M.A., Jouvent, E., MacIntosh, B.J., et al., 2019. Harmonizing brain magnetic resonance imaging methods for vascular contributions to neurodegeneration. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 11, 191–204.

Sudre, C.H., Anson, B.G., Ingala, S., Lane, C.D., Jimenez, D., Haider, L., Varsavsky, T., Tanno, R., Smith, L., Ourselin, S., et al., 2019. Let's agree to disagree: Learning highly debatable

multirater labelling, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 665–673.

Sudre, C.H., Cardoso, M.J., Bouvy, W.H., Biessels, G.J., Barnes, J., Ourselin, S., 2015. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. IEEE transactions on medical imaging 34, 2079–2102.

Tillin, T., Hughes, A.D., Mayet, J., Whincup, P., Sattar, N., Forouhi, N.G., McKeigue, P.M., Chaturvedi, N., 2013. The relationship between metabolic risk factors and incident cardiovascular disease in europeans, south asians, and african caribbeans: Sabre (southall and brent revisited)—a prospective population-based study. Journal of the American College of Cardiology 61, 1777–1786.

Timmins, K.M., van der Schaaf, I.C., Bennink, E., Ruigrok, Y.M., An, X., Baumgartner, M., Bourdon, P., De Feo, R., Noto, T.D., Dubost, F., Fava-Sanches, A., Feng, X., Giroud, C., Group, I., Hu, M., Jaeger, P.F., Kaiponen, J., Klimont, M., Li, Y., Li, H., Lin, Y., Loehr, T., Ma, J., Maier-Hein, K.H., Marie, G., Menze, B., Richiardi, J., Rjiba, S., Shah, D., Shit, S., Tohka, J., Urruty, T., Walińska, U., Yang, X., Yang, Y., Yin, Y., Velthuis, B.K., Kuijf, H.J., 2021. Comparing methods of detecting and segmenting unruptured intracranial aneurysms on tof-mras: The adam challenge. NeuroImage 238, 118216. URL: https://www.sciencedirect.com/science/article/pii/S1053811921004936, doi:https://doi.org/10.1016/j.neuroimage.2021.118216.

Van Straaten, E.C., Fazekas, F., Rostrup, E., Scheltens, P., Schmidt, R., Pantoni, L., Inzitari, D., Waldemar, G., Erkinjuntti, T., Mäntylä, R., et al., 2006. Impact of white matter hyperintensities scoring method on correlations with clinical data: the ladis study. Stroke 37, 836–840.

Vernooij, M., van der Lugt, A., Ikram, M.A., Wielopolski, P., Niessen, W., Hofman, A., Krestin, G., Breteler, M., 2008. Prevalence and risk factors of cerebral microbleeds: the rotterdam scan study. Neurology 70, 1208–1214.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., T O'Brien, J., Barkhof, F., Benavente, O.R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology 12, 822–838.

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Scientific reports 11, 1–15.

Wright, L., Demeure, N., 2021. Ranger21: a synergistic deep learning optimizer. arXiv preprint arXiv:2106.13731 .

Yates, P.A., Villemagne, V.L., Ellis, K.A., Desmond, P.M., Masters, C.L., Rowe, C.C., 2014. Cerebral microbleeds: a review of clinical, genetic, and neuroimaging associations. Frontiers in neurology 4, 205.

Yushkevich, P.A., Gao, Y., Gerig, G., 2016. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 3342–3345.

Zhang, C., Chen, Q., Wang, Y., Zhao, X., Wang, C., Liu, L., Pu, Y., Zou, X., Du, W., Pan, Y., et al., 2014. Risk factors of dilated virchow-robin spaces are different in various brain regions. PloS one 9, e105505.

Zhu, Y.C., Tzourio, C., Soumaré, A., Mazoyer, B., Dufouil, C., Chabriat, H., 2010. Severity of dilated virchow-robin spaces is associated with age, blood pressure, and mri markers of small vessel disease: a population-based study. Stroke 41, 2483–2490.