Original Article

JCBFM

Journal of Cerebral Blood Flow & Metabolism
0(00) 1–10

DOI: 10.1177/0271678X15610586
jcbfm.sagepub.com

SAGE

A benchmarking tool to evaluate computer tomography perfusion infarct core predictions against a DWI standard

Carlo W Cereda[1,2], Søren Christensen[1], Bruce CV Campbell[3,4], Nishant K Mishra[1], Michael Mlynash[1], Christopher Levi[5], Matus Straka[1], Max Wintermark[1], Roland Bammer[1], Gregory W Albers[1], Mark W Parsons[5] and Maarten G Lansberg[1]

Abstract

Differences in research methodology have hampered the optimization of Computer Tomography Perfusion (CTP) for identification of the ischemic core. We aim to optimize CTP core identification using a novel benchmarking tool. The benchmarking tool consists of an imaging library and a statistical analysis algorithm to evaluate the performance of CTP. The tool was used to optimize and evaluate an in-house developed CTP-software algorithm. Imaging data of 103 acute stroke patients were included in the benchmarking tool. Median time from stroke onset to CT was 185 min (IQR 180-238), and the median time between completion of CT and start of MRI was 36 min (IQR 25-79). Volumetric accuracy of the CTP-ROIs was optimal at an rCBF threshold of <38%; at this threshold, the mean difference was 0.3 ml (SD 19.8 ml), the mean absolute difference was 14.3 (SD 13.7) ml, and CTP was 67% sensitive and 87% specific for identification of DWI positive tissue voxels. The benchmarking tool can play an important role in optimizing CTP software as it provides investigators with a novel method to directly compare the performance of alternative CTP software packages.

Keywords

Cerebrovascular disease, brain imaging, brain ischemia, cerebral blood flow measurement, diffusion weighted MRI

Received 6 May 2015; Revised 20 August 2015; Accepted 27 August 2015

Introduction

Several studies have identified CTP parameters that could serve as surrogates for DWI imaging. These studies have, however, reported different results in terms of the optimal perfusion parameter (e.g. CBF vs CBV) and its optimal threshold to identify the ischemic core. This variability is attributed to differences in (1) CTP processing algorithms, (2) definitions of the gold standard for ischemic core, and (3) implementations of ROC analysis (definition of true negative region, ROC analysis per-patient or all voxels pooled)[1–9] (Table 1). Consequently, there is wide variability in the CTP parameters that are used in clinical practice and trials. This presents a major obstacle to progress in the field of CTP-based patient selection for acute stroke therapy. To address the heterogeneity of prior studies, we developed a benchmarking tool that can be used to evaluate CTP post-processing software algorithms in a standardized way. We used this tool to evaluate the performance of in-house developed CTP post-processing software algorithms.

[1]Stanford Stroke Center, Stanford University Medical Center, Stanford, CA, USA
[2]Stroke Center, Neurocenter (EOC) of Southern Switzerland, Lugano, Switzerland
[3]Departments of Medicine and Neurology, Melbourne Brain Centre at the Royal Melbourne Hospital, University of Melbourne, Parkville, Australia
[4]Department of Radiology, Royal Melbourne Hospital, University of Melbourne, Parkville, Australia
[5]Department of Neurology, John Hunter Hospital, University of Newcastle and Hunter Medical Research Institute, Newcastle, Australia

Corresponding author:
MG Lansberg, Stanford Stroke Center, 780 Welch Road, Suite 350, Palo Alto, CA 94304-5778, USA.
Email: lansberg@stanford.edu

Table 1. Summary of studies comparing CTP-based infarct prediction to a DWI reference.

Study	*N*	*CT-MRI interval Inclusion criteria median [IQR]*	*DWI outline method*	*Reference region for ROC analysis*	*CTP parameters tested*	*Optimal CTP threshold for core prediction*	*Volume difference DWI-CTP, ml Mean (95% CI)*
Current Study	103	≤3 h 36 min [25–77]	Manual	Hypo-perfused region (T_{max}>4 s)	rCBF	<38%	0.3 (−3.6 to 4.2)
Wintermark [8]	25	≤12 h 25 min [–]	Manual	Whole brain	CBV rCBV MTT/rMTT TTP/rTTP CBF/rCBF	<2.0 ml/100 g <60%	–
Bivard [3]	57 [a]	– 28 min [20–40]	Semi-automatic	Ipsilesional hemisphere	rCBF CBV/rCBV T_{max} MTT	<45%	–
Bivard [2]	67 [a]	≤12 h 162 min [185–240]	Manual	Ipsilesional hemisphere	rCBF CBF MTT TTP/rTTP CBV/rCBV	<40%	–
Kamalian[7]	48	≤1 h 34 min [28–43]	Semi-automatic	Not specified	rCBF CBF CBV/rCBV MTT/rMTT	<16–32% depending on software used	–
Campbell[4,6]	54 [a]	≤1 h 27 min [25–35]	Manual	Hypo-perfused region (rTTP>2 s)	rCBF CBF CBV/rCBV MTT T_{max} TTP	<31%	11.0 (–)
Bivard[1]	67 [a]	≤1 h – [–]	Semi-automatic	Ipsilesional hemisphere	rCBF CBF CBV/rCBV T_{max}/DT	<40%	0.5 (−0.1 to 0.9)
Bivard[5]	33 [a]	≤1 h – [–]	manual	Ipsilesional hemisphere	rCBF CBF CBV/rCBV MTT TTP	<50%	4.2 (3.4 to 7.2)
Schaefer[9]	55	≤3 h 51 min [41–65]	Semi-automatic	Ipsilesional hemisphere	rCBF CBV	<15%	−2.6 (–)[b]

– = not provided; IQR = interquartile range; ROC = Receiver operating characteristic; CTP = computer tomography perfusion; DWI = diffusion weighted imaging; CI = confidence interval; R^2 = correlation coefficient; h = hours; CBV = cerebral blood volume; rCBV = relative cerebral blood volume; MTT = ; rMTT = relative mean transit time; TTP = time to peak; rTTP = relative time to peak; CBF = cerebral blood flow, rCBF = relative cerebral blood flow; T_{max} = time-to-maximum of the residue function; DT = delay time. [a]Subset of these patients is included in the current study. [b]95% CI not reported; prediction interval reported as ± 56.7.

Materials and methods

A schematic view of the benchmarking tool that was developed to evaluate CTP post-processing software algorithms is shown in Figure 1. To use the tool, investigators need to (1) generate CTP ischemic core masks (in DICOM format) by processing the included CTP source data (in DICOM format) with their own CTP post-processing software; (2) place their CTP ischemic core masks in a predefined folder structure; and (3) run the benchmarking tool's analysis executable program with the mask folder as input. The tool will then generate a performance report of the user's CTP post-processing software algorithm based on the correspondence between the CTP masks and the tool's included gold standard DWI lesion masks using multiple metrics. Since the purpose of the tool is to provide an objective quantitative evaluation of the performance of CTP post-processing algorithms, all steps except for the perfusion algorithm itself, are standardized (Figure 1). In order to ensure the credibility and integrity of these steps, the tool features fully transparent and commented source code (Matlab v. R2013b, MathWorks Inc., Nattick, MA, USA) and a set of images for each case to verify the appropriateness of co-registration and DWI lesion outlines (Figure 2).

The benchmarking tool has only two technical requirements of the CTP software that is evaluated, which ensures compatibility with all open-source and most commercial CTP software packages: (1) the CTP software should output infarct mask data in the same pixel dimensions as the provided CTP input data (256×256 matrix); (2) the CTP software should not perform motion correction or spatial down-sampling because the CTP input data has already been motion corrected to ensure spatial correspondence with the co-registered DWI lesion outlines. The two main components of the benchmarking tool are:

1. **A large multicenter imaging dataset** from acute stroke patients who underwent back-to-back CTP and DWI imaging within 3 h of each other. Imaging data from two prospective cohort studies of acute ischemic stroke patients were pooled.[10,11] Imaging was performed at three US sites and one Australian site with CT and MRI scanners from all the major manufacturers. CT perfusion acquisition modes included toggle table, continuous spiral and cine mode with total z-axis coverage ranging from 4.4 to 16 cm. Tube voltage was constant at 80 kV across sites. Imaging data from patients who presented within 8 h of stroke onset and underwent an MRI within 3 h after CT were included for use in the benchmarking tool.

For each case, the acute DWI image was co-registered to the CTP slab(s) using the non-contrast CT as an intermediary target (Figure 2). The DWI image was then resampled to match the CTP slab and visually checked for accurateness using interactive image

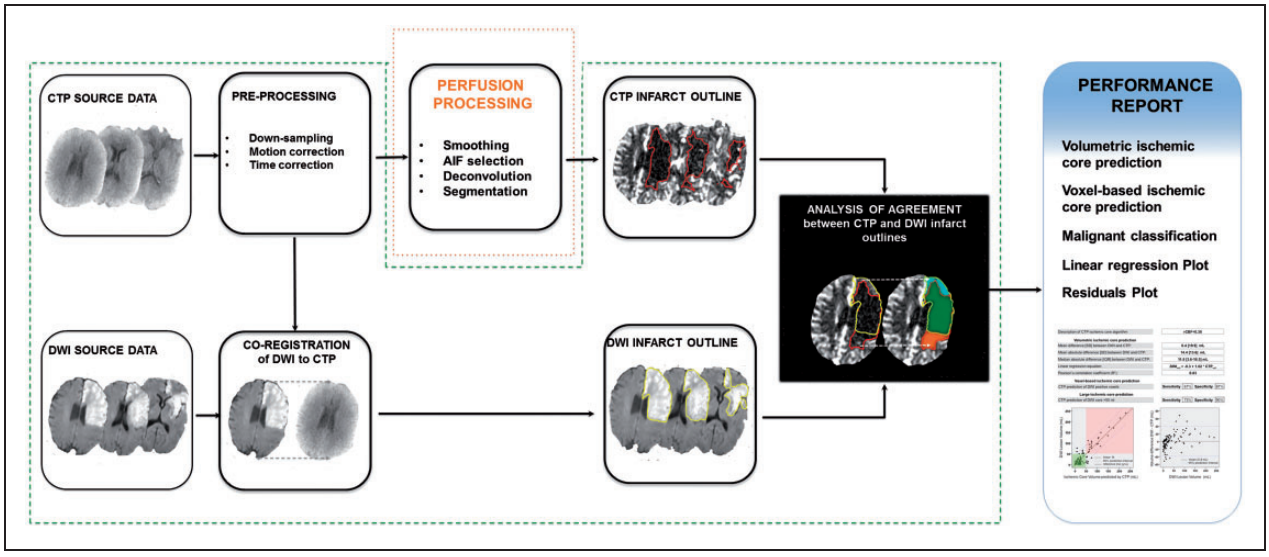

Figure 1. Flowchart of the processing steps required to compare CTP-defined ischemic core to a DWI standard reference. Each processing step can impact the perfusion analysis and influence the observed performance of CTP in terms of infarct core prediction. Only studies that use equivalent implementations are directly comparable. To allow for better inter-study comparability, we have created an open-source CTP benchmarking tool in which all processes are standardized (green outline: co-registration, DWI lesion outlining, and statistical evaluation), except for the CTP algorithm used to segment the infarct core (orange outline).

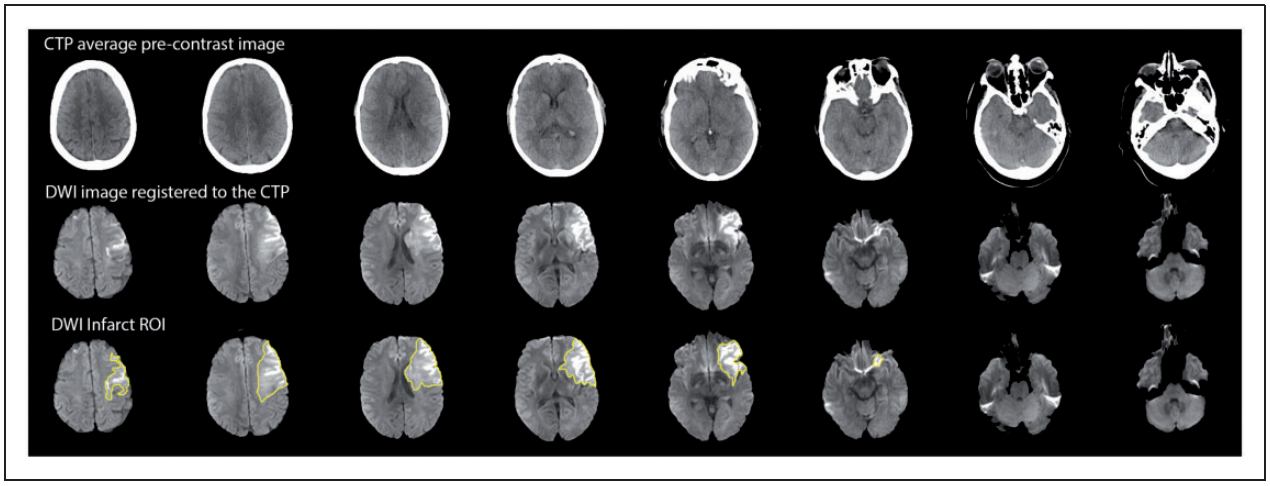

Figure 2. Example of co-registered CTP and DWI images along with the overlaid infarct ROI outline that form part of the benchmarking tool. These images allow users of this benchmarking tool to validate the appropriateness of the DWI outline as well as the DWI-to-CTP co-registration. For formatting purposes, the images are shown on three rows, but the actual format is three separate sets of images (per patient), which makes it easy to flick back and forth between images to assess the co-registration in a pixel wise fashion.

blending. Registrations were performed using MNI (Montreal Neurological Institute) tools and subject to quality verification and approval by three investigators.[12] DWI lesion ROIs were drawn on the DWI images following resampling to CTP space by a single investigator (BC) and subjected to group review until all outlines were accepted. This procedure was fully blinded to the CTP maps. Figure 3 details the regions used in our analyses. Tissue with normal perfusion (Tmax$\leq$4 s), such as the contralateral hemisphere, was excluded. Cases in which more than 50% of the DWI lesion had normal perfusion at the time of CTP were excluded from the study on the grounds of major hemodynamic changes.[4]

2. An open source **statistical analysis program** to quantify the correspondence between CTP and DWI imaging in a reproducible way. The program expresses performance of the CTP software for identification of the ischemic core using five different metrics: (1) mean prediction error (mean difference between the CTP and DWI ischemic core volumes); (2) mean absolute prediction error; (3) regression and correlation coefficients of the relationship between the CTP and DWI ischemic core volumes; (4) sensitivity and specificity of CTP to identify DWI positive voxels; and (5) sensitivity and specificity of CTP to identify patients with a DWI core volume exceeding 50 ml. The program displays performance visually on a scatter plot with CTP volumes on the *x*-axis and DWI volumes on the *y*-axis and on a residual plot with DWI volumes on the *x*-axis and the difference between the CTP and DWI volumes on the *y*-axis.

The benchmarking tool was used to assess the performance of a research version of an in-house developed, fully automated, CTP post-processing software algorithm.[13] This algorithm identifies segments of tissue with a relative CBF ($rCBF = CBF_{voxel}/CBF_{control}$) below a configurable threshold, where $CBF_{control}$ is defined as the mean CBF of tissue with normal perfusion. Our CTP software algorithm was first run with its default rCBF threshold (rCBF $<$30%) to generate ischemic core segmentation masks for each case. Next, the CTP software was set up to produce 27 segmentation masks of the ischemic core per case. This was based on 27 rCBF thresholds ranging from 0 to 1 with the finest resolution (0.02) between 0.2 and 0.5 as prior studies and previous experience with our perfusion algorithm indicated this range to be the most relevant for segmentation of the ischemic core.[4] Three optimal rCBF thresholds were determined: (1) a volume-optimized rCBF threshold defined as the threshold at which the mean difference between predicted core volumes and observed DWI volumes was minimized[4], (2) a volume-optimized rCBF threshold defined as the threshold at which the median absolute difference between predicted core volumes and observed DWI volumes was minimized and (3) a voxel-optimized rCBF threshold defined as the threshold at which the Youden's index based on ROC analysis for predicting DWI positive voxels was maximized.[14] The three sets of lesion masks,

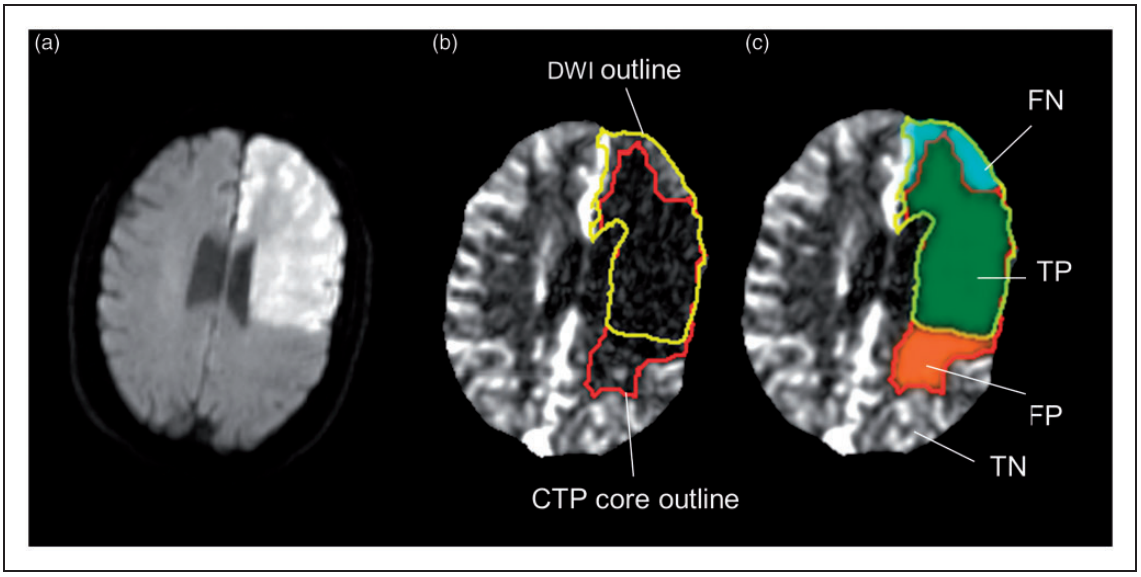

Figure 3. Illustration of voxel-based analysis of infarct prediction. This case illustrates how the CTP benchmarking tool calculates the test-characteristics of CTP for identifying DWI positive voxels. The DWI (a) is co-registered to the CTP (b and c). The rCBF estimate (red outline) of the ischemic core and the gold standard DWI (yellow) are shown in panel (b). CTP test characteristics are based on the regions shown in panel (c). TP (true positive) are voxels that are included in the CTP-rCBF and the DWI infarct outlines (green); FP (false positive) are voxels included in the CTP-rCBF outline but not the DWI outline (red); FN (false negative) are voxels included in the DWI outline but not the CTP-rCBF outline (blue); TN (true negative) are voxels that are not included in the DWI or the CTP-rCBF outline but have prolonged Tmax defined as Tmax>4 (region not shown to maintain a clear depiction of the rCBF and DWI outlines).

generated with these rCBF thresholds, were used as input for the benchmarking tool's analysis program to generate performance reports.

We conducted analyses to determine if the volume-optimized rCBF threshold varied with (1) the time from symptom onset to CTP and (2) the time from CTP to MRI. For each individual patient, we determined a patient-specific optimal rCBF threshold (based on minimal absolute volumetric difference between CTP and DWI core volumes). These patient-specific optimal rCBF thresholds were regressed, separately, against the onset-to-CT and CT-to-MRI time intervals to determine if significant associations existed. We also performed a sensitivity analysis to determine if the optimal rCBF threshold differed depending on the *z*-axis coverage of the CTP scan (4.8 vs 8 vs 16 cm).

Finally, we repeated the optimization procedure for a perfusion algorithm that uses rCBV for identification of the ischemic core and compared the prediction errors of the two approaches (rCBV and rCBF based) with a paired Wilcoxon signed rank test.

Results

The pooled dataset included 128 patients who were enrolled in the parent studies between 2004 and 2012 and who had undergone back-to-back CTP and diffusion MRI in the acute stroke setting. Of these, 103 patients met eligibility criteria for this study. Patients were excluded because more than 50% of the DWI lesion had normal perfusion at the time of CTP ($n = 18$), there was insufficient quality of the baseline CTP data ($n = 4$) and coregistration failures due to image distortions ($n = 3$). The mean age of the included population was 68 years (SD 14), median baseline National Institutes of Health Stroke Scale (NIHSS) score was 16 (IQR 11-19), median time from stroke onset to CT was 185 min (IQR 180-238) and the median time between completion of CT and start of MR was 36 min (IQR 25-79, range 15–181 min). Twenty-nine patients received intravenous thrombolysis only, 16 underwent endovascular therapy only, 14 had both therapies and the remaining 44 had no revascularization therapy.

The rCBF threshold that optimized the mean difference between CTP and DWI lesion volumes was 38% (Figures 4 and 5a). The rCBF threshold that optimized the median absolute difference between CTP and DWI lesion volumes was 30% (Supplemental Figure). The rCBF threshold at which the Youden's index was maximized was 42%. (Youden's index 0.55; Figure 5b) The software's performance characteristics at these thresholds are listed in Table 2.

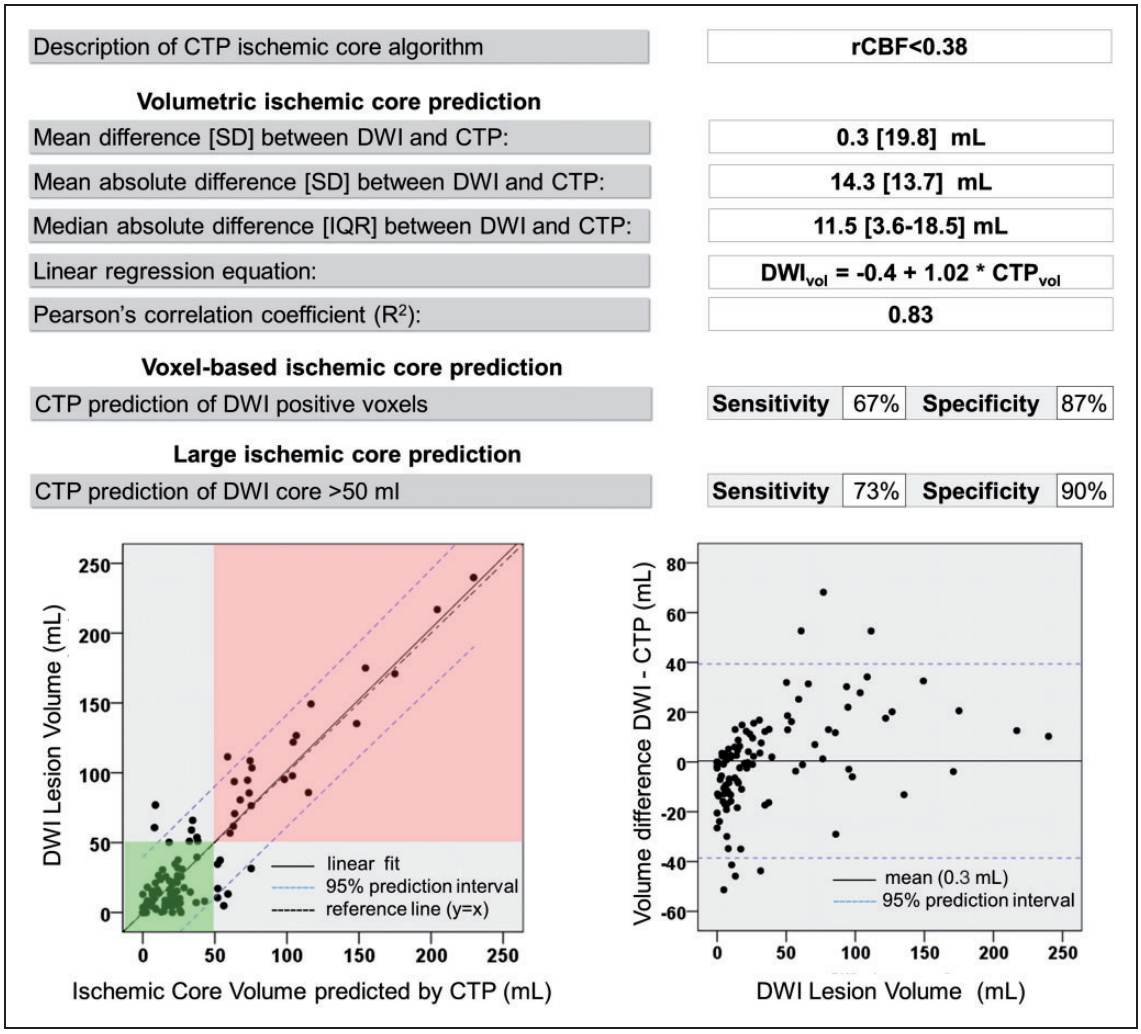

Figure 4. Example performance report of CTP software operating at rCBF<38% threshold. This report card, generated with the CTP benchmarking tool, lists the performance metrics of our in-house CTP analysis software operating at its volume-optimized rCBF threshold of <38%. Lower left graph: A scatter plot of CTP and DWI ischemic core lesion volumes is shown with a linear regression line (black) and its 95% prediction interval between blue dashed lines. The green shaded area indicates patients who are correctly classified by CTP as having a DWI lesion <50 ml. The red shaded area indicates patients who are correctly classified by CTP as having a DWI lesion >50 ml. Lower right graph: A residuals plot shows the volumetric difference between CTP infarct core prediction and DWI. The mean difference between the DWI and CTP infarct volumes (mean error calculated as DWIvol − CTPvol) is indicated with a black line and its 95% prediction interval with blue dashed lines.

There was no significant association between patient-specific optimal rCBF thresholds and the onset-to-CT interval ($R^2 = 0.003$; $p = 0.56$) or the CT-to-MRI interval ($R^2 = 0.005$; $p = 0.48$). Sensitivity analyses in subgroups defined by CTP z-axis coverage (4.4 vs 8 vs 16 cm) also yielded identical volume-optimized rCBF thresholds (<38%).

The optimal rCBV threshold for prediction of the DWI core was <44%. The performance characteristics for ischemic core segmentation were similar with the optimal rCBV (<44%) and rCBF (<38%) method (Table 2). The mean difference in prediction errors between methods was 0.2 ml (SD 14.0 ml; $p = 0.55$).

Discussion

We developed a benchmarking tool that standardizes the evaluation of CTP software for ischemic core prediction and we used this tool to evaluate in-house developed CTP software. At the optimal rCBF threshold (<38%), the mean absolute difference between ischemic core lesion volumes assessed with CTP and DWI was 14.3 ml with a standard deviation of 13.7 ml. This result can serve as an initial benchmark for the performance of other CTP software packages.

A novel aspect of the benchmarking tool is that it reports the mean absolute prediction error (difference

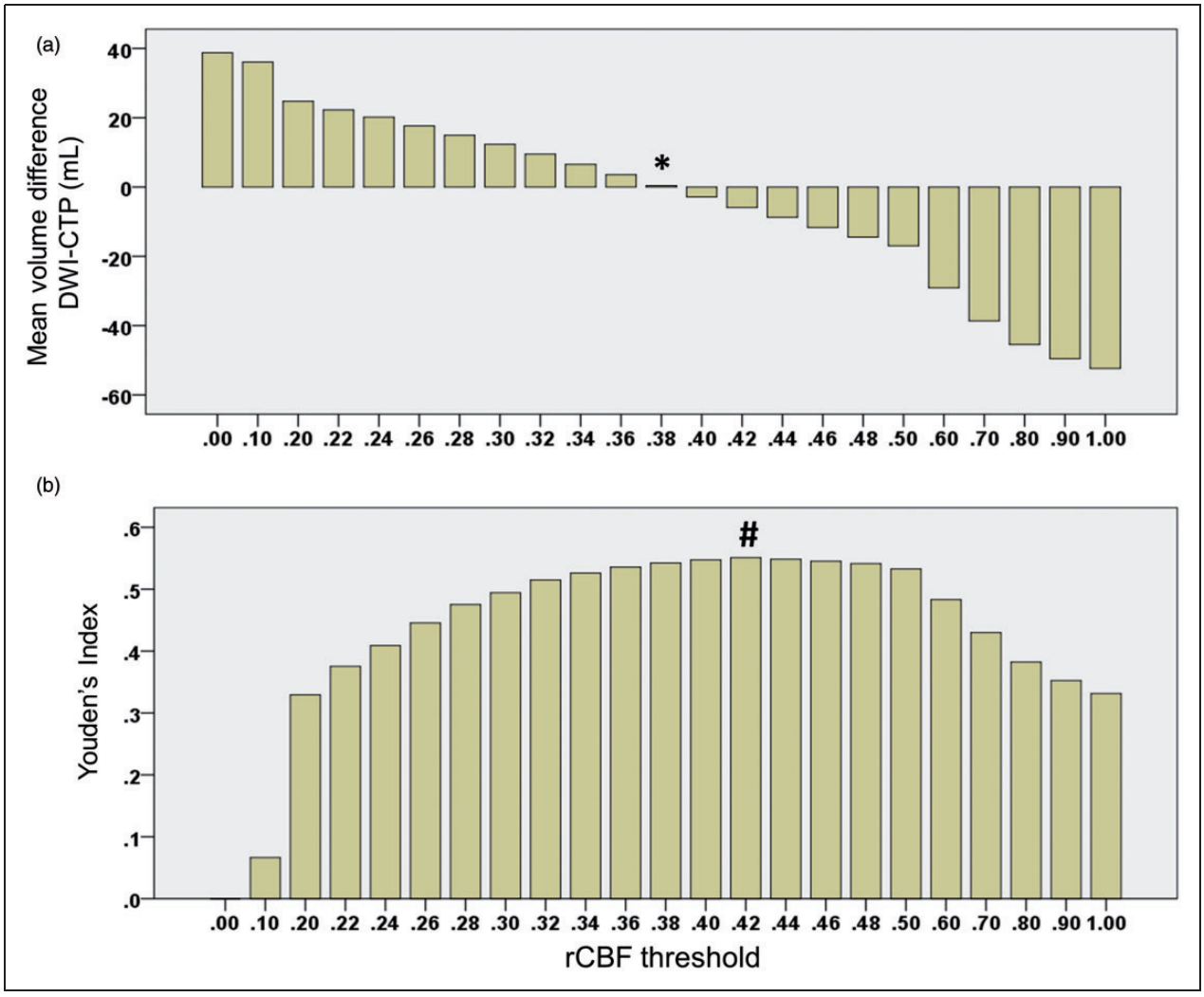

Figure 5. Optimal rCBF thresholds for DWI volume prediction. The volume-optimized threshold was defined as the rCBF threshold at which the mean prediction error (DWI-CTP infarct volume) was minimized. This occurred at an rCBF threshold <38%, which corresponded with a mean volumetric difference between DWI and CTP core volumes of 0.3 ml (panel A, column indicated by an asterisk). The ROC-optimized threshold was defined as the rCBF threshold at which the Youden's Index was maximal. This occurred at an rCBF threshold of <42%. At this threshold CTP was 72% sensitive and 83% specific for identifying DWI positive voxels, corresponding to a Youden's index of 0.55 (panel B, column indicated by a hash character).

between CTP and DWI lesion volumes). This measure has clinical relevance, as it reflects the accuracy with which ischemic core lesion volumes are measured using CTP at the level of individual patients. In contrast, the more commonly used mean prediction error is less informative as it only reflects the average bias between CTP and DWI volumes and does not provide information about the volumetric difference for an individual case. Other novel performance measures, calculated by the benchmarking tool, are the sensitivity and specificity of CTP for identifying patients with DWI lesions exceeding 50 ml (Table 2, Figure 4). The clinical relevance of these measures is based on recent studies that have shown poor outcome, regardless of reperfusion therapy, in patients with large DWI lesions and the exclusion of patients with large DWI lesions from recent endovascular stroke trials.[15,16]

Another novel aspect of this research is the use of "open science". Whilst there has been enthusiasm to make scientific research methods, data and results publicly available,[17,18] such open science is still exceedingly rare. Limited sharing of research data and methods has hindered the advancement of scientific research, including research related to the optimization of CTP post-processing algorithms. Considerable variability in the quality and quantity of the CTP input data coupled with variability in the methods used to evaluate post-processing algorithms in prior studies has led to results that are inconsistent, impossible to replicate, and difficult to compare. Consequently, newer studies have not

Table 2. Performance characteristics of in-house developed CT perfusion software for identification of the infarct core.

	CTP threshold for ischemic core segmentation[a]			
Measure of agreement between CTP and DWI infarct core	rCBF <30%	rCBF <38%	rCBF <42%	rCBV <44%
Volumetric agreement (DWI – CTP lesion volume)				
Mean difference (SD), ml	12.0 (19.0)	0.3 (19.8)	−5.9 (21.3)	0.5 (18.1)
Mean absolute difference (SD), ml	15.8 (16.1)	14.3 (13.7)	16.2 (15.0)	12.9 (12.7)
Median absolute difference (IQR), ml	9.4 (4.6–22.3)	11.5 (3.6–18.5)	12.5 (5.6–22.3)	8.7 (4.1–19.5)
Regression intercept, coefficient ($DWI_{vol} = a + b \times CTP_{vol}$)	7.9, 1.16	−0.4, 1.02	−4.2, 0.96	1.4, 0.96
Pearson Correlation (R^2)	0.86	0.83	0.81	0.86
Spatial agreement				
Sensitivity of CTP for predicting DWI positive voxels	55%	67%	72%	69%
Specificity of CTP for predicting DWI positive voxels	95%	87%	83%	88%
Agreement for identification of large infarct core				
Accuracy of CTP for predicting DWI core exceeding 50 ml	87%	85%	86%	89%
Sensitivity of CTP for predicting DWI core exceeding 50 ml	60%	73%	77%	77%
Specificity of CTP for predicting DWI core exceeding 50 ml	99%	90%	90%	95%

IQR = interquartile range; DWI_{vol} = infarct core volume estimated by DWI; CTP_{vol} = infarct core volume estimated by CTP; R^2 = correlation coefficient squared. [a]rCBF<30% is the default threshold of our CTP software and the rCBF threshold at which the median absolute difference between the DWI and CTP infarct volumes is minimized; rCBF<38% is the mean volume-optimized rCBF threshold, defined as the rCBF threshold at which the mean difference between the DWI and CTP infarct volumes is minimized; rCBF<42% is the voxel-optimized rCBF threshold, defined as the rCBF threshold at which the Youden's index based on ROC analysis for predicting DWI positive voxels is maximized; rCBV<44% is the volume-optimized rCBV threshold, defined as the rCBV threshold at which the mean difference between the DWI and CTP infarct volumes is minimized.

been able to build on the results of older studies to incrementally improve CTP post-processing algorithms, there has been no clear movement towards a consensus among scientists of what constitutes an adequate quality for CTP post-processing algorithms, and there is considerable variability in the type of algorithms that are being used. By making our imaging data and evaluation methods available to others, we aim to create a global research environment that is conducive to continuous improvements of CTP post-processing software algorithms.

Our CTP post-processing algorithm showed the smallest difference between CTP and DWI ischemic core estimates at an rCBF threshold <38%. This threshold is in the range of rCBF thresholds suggested in prior studies (<31% to <50%) (Table 1). The spread in rCBF thresholds among studies is likely due to the wide variety of methodological approaches (acquisition, post-processing and analysis strategies) used in these studies. For example, our analyses illustrate how the choice of the optimization parameter impacts the rCBF threshold. The rCBF threshold was <38% when optimized for absolute volumetric correspondence and <42% when a voxel-based optimization was employed.

Voxel-based optimization has several limitations. First, voxel-based optimization does not guarantee volumetric agreement. In our case, the higher, more sensitive, rCBF threshold identified with voxel-based optimization (<42%) results in an overestimation of the ischemic core volume compared to DWI (mean overestimation of 6 ml). Second, voxel-based optimization depends on the region in which it is assessed (e.g. whole brain, ipsilesional brain, hypo-perfused region). The choice of reference region is an arbitrary decision that varies between studies. Finally, voxel-based optimization is more sensitive to co-registration errors than a volume-based approach. While we took extreme care to optimize co-registration in the imaging dataset, minor errors are unavoidable because of the many challenges of registering between CTP and DWI modalities, including non-isotropic data, different slice angulations, and inherent distortions in DWI images. The focus on sensitivity and specificity in prior studies could have been motivated by the fact that good volumetric correspondence does not necessarily imply good spatial correspondence in the individual patient. This indeed may be a concern for small samples, but when the CTP algorithm is evaluated in a large patient sample, like the one used for this study, good volumetric agreement implies good spatial agreement. For these reasons, we favor volumetric optimization, complemented by ROC and visual analysis to summarize and ensure acceptable spatial concordance.

From a clinical standpoint, a threshold that is more restrictive and thus more specific than the threshold at which the mean volumetric difference is optimized may be desirable, because a more restrictive threshold would err on the side of underestimating the ischemic core. This avoids falsely identifying a patient as a poor

candidate for reperfusion therapy based on a large ischemic core by CTP, when he or she would have been considered a good candidate based on a smaller DWI lesion. For example, the rCBF<30% threshold, on average, underestimates the DWI lesion by 12 ml; however, it has greater specificity for predicting DWI positive voxels compared to the rCBF<38% threshold (95% vs 87%). Because of the bias towards underestimation, the rCBF<30% threshold is less likely to overestimate the ischemic core than the <38% threshold and when overestimation occurs, the volume by which it overestimates is smaller.

In this study, we demonstrate similar prediction errors with rCBF and rCBV based segmentation. We foresee that innovations in CTP post-processing algorithms will reduce CTP prediction errors. These innovations could focus on improvements in rCBF or rCBV thresholds or may, instead, be based on alternative perfusion parameters such as T_{max}, used alone or in combination with CBF and/or CBV criteria. Another potential area of improvement is the use of different thresholds for gray and white matter. The single threshold used in this study and in most prior studies tends to overestimate the lesion in white matter while underestimating it in gray matter. In our dataset, this results in overestimation of small DWI lesions (<15 ml) and underestimation of larger DWI lesions. The benchmarking tool is ideal for testing the performance of novel algorithms, as the tool is not specific to rCBF or rCBV segmentation, but can evaluate the performance of any CTP post-processing software regardless of its segmentation algorithm.

A limitation of the benchmarking tool is the lack of an ideal gold standard for ischemic core. Two MRI sequences have traditionally been used: (1) DWI obtained early after CTP and (2) FLAIR/T2 obtained at late follow-up in patients with documented early reperfusion. Both methods are imperfect since the ischemic core is expected to expand between the time of CTP imaging and DWI as well as between CTP imaging and reperfusion. Consequently, the MRI gold standard overestimates the true ischemic core volume with both approaches. To address this limitation, we only included patients in whom the CTP and MRI were obtained back-to-back (median time delay 35 min), but even in this short time-frame some ischemic core growth may occur. This can affect the estimate of the optimal rCBF threshold, as longer CT-to-MRI intervals would be associated with higher rCBF thresholds (and consequently larger CTP cores) to compensate for greater overestimation of the ischemic core on DWI. We, however, found no association between time from CT-to-MRI and the optimal rCBF threshold, suggesting that the dataset is sufficiently uniform in terms of the patients' CT-to-MRI intervals. Additional factors that make DWI an imperfect gold standard include the effects of edema, partial reversal of the DWI after reperfusion,[19,20] and the inherently subjective nature of lesion outlines.[21] Consequently, MRI can over- or underestimate the ischemic core in individual patients, which makes it unrealistic to expect perfect concordance between ischemic core measurements on CT and MRI.

A second limitation is the potential dependency of the optimal CTP threshold on the duration between symptom onset and CTP (ie onset-to-CT time). Fundamentally, infarction is expected to depend on the duration and the severity of CBF reduction, with a lower (more restrictive) rCBF threshold required for patients who are scanned early (i.e. short duration of ischemia) and a higher threshold for patients scanned late. In our dataset, we detected no dependency of the optimal rCBF threshold on the onset-to-CT time. This is consistent with other studies. It suggests that using our approach in the time-window studied, the time-dependency of the core rCBF threshold could not be documented.[22]

We foresee that the CTP benchmarking tool will evolve over time as researchers add functionality. A current strength of the tool is the generalizability of its results. This stems from the large dataset of back-to-back CTP-DWI images, obtained with a wide array of acquisition protocols on CT scanners from all major vendors at multiple sites. Nevertheless, cases could be added to cover, for example, a wider range of lesion volumes. Addition of follow-up imaging and data on reperfusion could allow development of a complimentary module that evaluates CTP-based segmentation of critically hypo-perfused tissue. Finally, alternative measures of performance, such as for example the mean squared error or DICE coefficient, can easily be added.

The benchmarking tool that we developed will provide a transparent platform for comparison of CTP ischemic core segmentation algorithms and has the potential to play a major role in advancing the diagnostic accuracy of CTP software. It demonstrates that ischemic core volumes predicted by our in-house developed CTP software differ, on average, by 14.3 ml (SD 13.7) from DWI core volumes when using an rCBF threshold of <38%. This should be viewed as an initial benchmark, and we anticipate that future efforts will lead to better algorithms that generate CTP lesion volumes that approximate the DWI volumes even more closely.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article:

The study was funded by grants from the National Institute for Neurological Disorders and Stroke (NINDS).

R01 NS03932505 (G. Albers), 5 R01 NS075209 (M. Lansberg), and Medical Research Council partnership project grant (ID: 1013719) (M. Parsons).

Authors' contributions

The study was designed by MGL, GA, SC, CWC, BC, MS, and RB. Data were collected MGL, GA, SC, BC, MP, and CL. Data were analyzed and interpreted by MGL, CWC, SC, BC, NM, MM, MP, and GA. The manuscript was drafted by MGL, CWC, SC, BC, NM, MM, MP, and GA Critical revisions were made by all authors.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article:

G Albers has received consulting fees and expenses from Lundbeck for Steering Committee work and consulting fees from Concentric for serving on a Data Safely and Monitory Board. G Albers and R Bammer are equity shareholders in iSchemaView and perform consulting work for iSchemaView. Soren Christensen performs consulting work for iSchemaView.

Supplementary material

Supplementary material for this paper can be found at http://jcbfm.sagepub.com/content/by/supplemental-data

References

1. Bivard A, Levi C, Spratt N, et al. Perfusion ct in acute stroke: A comprehensive analysis of infarct and penumbra. *Radiology* 2013; 267: 543–550.
2. Bivard A, Spratt N, Levi C, et al. Perfusion computer tomography: Imaging and clinical validation in acute ischaemic stroke. *Brain: J Neurol* 2011; 134: 3408–3416.
3. Bivard A, McElduff P, Spratt N, et al. Defining the extent of irreversible brain ischemia using perfusion computed tomography. *Cerebrovasc Dis* 2011; 31: 238–245.
4. Campbell BC, Christensen S, Levi CR, et al. Cerebral blood flow is the optimal ct perfusion parameter for assessing infarct core. *Stroke* 2011; 42: 3435–3440.
5. Bivard A, Levi C, Krishnamurthy V, et al. Defining acute ischemic stroke tissue pathophysiology with whole brain ct perfusion. *J Neuroradiol* 2014; 41: 307–315.
6. Campbell BC, Christensen S, Levi CR, et al. Comparison of computed tomography perfusion and magnetic resonance imaging perfusion-diffusion mismatch in ischemic stroke. *Stroke* 2012; 43: 2648–2653.
7. Kamalian S, Kamalian S, Maas MB, et al. Ct cerebral blood flow maps optimally correlate with admission diffusion-weighted imaging in acute stroke but thresholds vary by postprocessing platform. *Stroke* 2011; 42: 1923–1928.
8. Wintermark M, Flanders AE, Velthuis B, et al. Perfusion-ct assessment of infarct core and penumbra: Receiver operating characteristic curve analysis in 130 patients suspected of acute hemispheric stroke. *Stroke* 2006; 37: 979–985.
9. Schaefer PW, Souza L, Kamalian S, et al. Limited reliability of computed tomographic perfusion acute infarct volume measurements compared with diffusion-weighted imaging in anterior circulation stroke. *Stroke* 2015; 46: 419–424.
10. Lansberg MG, Straka M, Kemp S, et al. Mri profile and response to endovascular reperfusion after stroke (defuse 2): A prospective cohort study. *Lancet Neurol* 2012; 11: 860–867.
11. Lin L, Bivard A, Levi CR, et al. Comparison of computed tomographic and magnetic resonance perfusion measurements in acute ischemic stroke: Back-to-back quantitative analysis. *Stroke* 2014; 45: 1727–1732.
12. Minc, http://www.Bic.Mni.Mcgill.Ca/servicessoftware/minc (accessed 1 May 2015).
13. Straka M, Albers GW and Bammer R. Real-time diffusion-perfusion mismatch analysis in acute stroke. *J Magn Reson Imag* 2010; 32: 1024–1037.
14. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3: 32–35.
15. Inoue M, Mlynash M, Straka M, et al. Patients with the malignant profile within 3 hours of symptom onset have very poor outcomes after intravenous tissue-type plasminogen activator therapy. *Stroke* 2012; 43: 2494–2496.
16. Saver JL, Goyal M, Bonafe A, et al. Stent-retriever thrombectomy after intravenous t-pa vs. T-pa alone in stroke. *N Engl J Med* 2015; 372: 2285–2295.
17. Poldrack RA and Gorgolewski KJ. Making big data open: Data sharing in neuroimaging. *Nat Neurosci* 2014; 17: 1510–1517.
18. Poline JB, Breeze JL, Ghosh S, et al. Data sharing in neuroimaging research. *Front Neuroinform* 2012; 6: 9.
19. Campbell B, Purushotham A, Christensen S, et al. The infarct core is well represented by the acute diffusion lesion: Sustained reversal is infrequent. *J Cerebral Blood Flow Metab: Official Journal of the International Society of Cerebral Blood Flow and Metabolism* 2012; 32: 50–56.
20. Soize S, Tisserand M, Charron S, et al. How sustained is 24-hour diffusion-weighted imaging lesion reversal? Serial magnetic resonance imaging in a patient cohort thrombolyzed within 4.5 hours of stroke onset. *Stroke* 2015; 46: 704–710.
21. Luby M, Bykowski JL, Schellinger PD, et al. Intra- and interrater reliability of ischemic lesion volume measurements on diffusion-weighted, mean transit time and fluid-attenuated inversion recovery mri. *Stroke* 2006; 37: 2951–2956.
22. Qiao Y, Zhu G, Patrie J, et al. Optimal perfusion computed tomographic thresholds for ischemic core and penumbra are not time dependent in the clinically relevant time window. *Stroke* 2014; 45: 1355–1362.