Accuracy and Performance of Functional Parameter Estimation Using a Novel Numerical Optimization Approach for GPU-Based Kinetic Compartmental Modeling

Igor Svistoun[1], Brandon Driscoll[1], and Catherine Coolens[1,2,3]

[1]Department of Medical Physics, Princess Margaret Cancer Centre and University Health Network, Toronto, Canada; [2]Departments of Radiation Oncology and IBBME, University of Toronto, Toronto, Canada; and [3]TECHNA Institute, University Health Network, Toronto, Canada

Corresponding Author:
Catherine Coolens, PhD
Princess Margaret Cancer Centre, Department of Medical Physics,
Rm 6:306 - 700 University Avenue, Toronto, ON M5G 1Z5, Canada;
E-mail: Catherine.Coolens@rmp.uhn.ca

Key Words: DCE imaging, numerical optimization, functional analysis, GPU

Abbreviations: Dynamic contrast-enhanced (DCE), graphical processing unit (GPU), magnetic resonance imaging (MRI), arterial input function (AIF), computer processing unit (CPU), finite impulse response (FIR), infinite impulse response (IIR), pattern search (PS), differential evolution (DE), fast Fourier transform (FFT)

ABSTRACT

Quantitative kinetic parameters derived from dynamic contrast-enhanced (DCE) data are dependent on signal measurement quality and choice of pharmacokinetic model. However, the fundamental optimization analysis method is equally important and its impact on pharmacokinetic parameters has been mostly overlooked. We examine the effects of those choices on accuracy and performance of parameter estimation using both computer processing unit and graphical processing unit (GPU) numerical optimization implementations and evaluate the improvements offered by a novel optimization approach. A test framework was developed where experimentally derived population-average arterial input function and randomly sampled parameter sets {K_{trans}, K_{ep}, V_b, τ} were used to generate known tissue curves. Five numerical optimization algorithms were evaluated: sequential quadratic programming, downhill simplex (Nelder–Mead), pattern search, simulated annealing, and differential evolution. This was combined with various objective function implementation details: delay approximation, discretization and varying sampling rates. Then, impact of noise and CPU/GPU implementation was tested for speed and accuracy. Finally, the optimal method was compared to conventional implementation as applied to clinical DCE computed tomography. Nelder–Mead, differential evolution and sequential quadratic programming produced good results on clean and noisy input data outperforming simulated annealing and pattern search in terms of speed and accuracy in the respective order of 10^{-8}%, 10^{-7}%, and $\times 10^{-6}$%). A novel approach for DCE numerical optimization (infinite impulse response with fractional delay approximation) was implemented on GPU for speed increase of at least 2 orders of magnitude. Applied to clinical data, the magnitude of overall parameter error was $<$10%.

INTRODUCTION

Obtaining a better understanding of a (personalized) tumor or disease microenvironment is quickly becoming a driving force in a whole range of medical scenarios from earlier disease diagnosis to image-based assessment of treatment efficacy (1). In this context, dynamic contrast-enhanced (DCE) imaging is increasingly used to help quantify vascular and tissue properties as to inform on the functionality and dynamic behavior of the disease and/or normal tissue. In terms of tissue perfusion and permeability, this is typically achieved with the additional use of tracer kinetic models that describe the flow of contrast agents through the tissue (2).

DCE computed tomography (CT) and magnetic resonance imaging (MRI) have been widely investigated, and despite their obvious differences in methodology to measure dynamic contrast enhancement curves, they share the same parametric analysis approach: both use low-molecular-weight contrast agents and as such they share mostly the same pharmacokinetic models that are applied after the imaging signal is converted to contrast concentration data (3). The delivery of the contrast agent to the organ or region of interest (eg, a tumor) is reflected in the arterial input function (AIF). Using the contrast enhancement curves in the organ or region of interest as a response on the AIF, an estimation of the tracer kinetic model parameters can be obtained. An example of this would be the widely used 2-compartmental modified Tofts model (2).

Whereas increasing efforts are in place to help standardize the acquisition and analysis methods of DCE imaging in both CT

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00048

Table 1. Tofts Model Parameters

Variable	Description	Units
C_t	Tissue concentration of contrast agent as a function of time	HU
C_a	AIF representing the arterial concentration of contrast agent as a function of time	HU
K_{trans}	Transfer constant from blood plasma into the EES	mL/g/min
K_{ep}	Transfer constant from EES back to the blood plasma	mL/g/min
V_b	Blood volume per unit of tissue	mL/g
t	Time variable	second
τ	Time delay from time of contrast injection to contrast arriving at region of interest	second
HCT	Hematocrit—fraction of red blood cells in blood. Value of 0.4 is used during this investigation.	Fraction

(4) and MRI (5), the solution of these tracer kinetic models is not necessarily trivial and requires an optimization method to solve for parameters in heterogeneous volumetric data. The effect of image noise and voxel-based analysis has also been reported on, showing a marked improvement in parameter robustness that can be achieved by balancing preprocess filtering with information loss (6). Regardless, parameters must be extracted given the nonuniform, discrete, limited-time measurements. Implementing the parameter estimation algorithm involves many other design decisions including choice of data processing rate, continuous-to-discrete system mapping approach, and numerical optimization algorithm. To the best of our knowledge, no investigations have been reported on the impact of the optimization method used on resulting parametric maps. Yet, it is well-known from other areas of research that significant differences can be found in between optimization methods in their ability to adequately resolve multiple variables simultaneously.

Given the large amount of data involved in processing DCE parametric maps, it is further increasingly important that these processes are as automated as possible to allow for useful integration into clinical workflows with a nearly real-time experience. Current implementations of kinetic models rely on manual or semiautomated estimations of the fractional delay in contrast arrival time at the region of interest. Not only is this a time-limiting factor for a fully automated workflow, it will be shown that lack of inclusion of this parameter in the optimization process creates larger estimation errors. For this reason, moving the optimization processes to a graphical processing unit (GPU) offers known speed improvements over standard computer processing unit (CPU) implementation of a fully inclusive optimization approach.

Having recently shown the improved correlation between CT- and MRI-based perfusion parameters (7) when using a common analysis platform to process DCE data regardless of the imaging modality, the purpose of this paper is now to (1) quantify the effects of system design choices (eg, processing sampling rate) and noise (both aliasing and background) present in the data on accuracy and speed of various CPU and GPU numerical optimization implementations and (2) to obtain a better understanding of parameter accuracy in clinically relevant DCE-CT data.

METHODS

Continuous Time Model and Problem Statement

Various models exist to describe contrast solute exchange of iodine or gadolinium-based DCE imaging methods. The modified Tofts model is by far the most widely implemented technique and as such it was felt worthwhile to investigate the design variations to better understand the largest available literature of pharmacokinetic metrics reported. The modified Tofts model describes a linear time-invariant first-order system. Data are acquired by the scanner, which can be expressed as tissue concentration function $C_t[n]W_c[n]$ and AIF $C_a[n]W_c[n]$ for $n \in \{\text{nonuniform discrete time points}\}$. $W_c[n]$ is a rectangular window function that takes on the value 1 at $0 < n \leq c \,/\, T$ and 0 otherwise, where T is the sampling period. The window function represents the fact that acquisition of measurements stops after a certain time $= c$ seconds.

The 2-compartmental model of tissue enhancement that takes into account contributions from intravascular and the interstitial space (which is what's measured by the scanner) is given by the following linear time-invariant system (2):

$$\begin{aligned} C_t(t) &= C_a(t) \times \left[\left(\frac{K_{trans}}{1 - HCT} \right) e^{-K_{ep}(t-\tau)} u(t-\tau) + V_b \delta(t-\tau) \right] \\ &= C_a(t) \times H(t) \end{aligned} \quad (1)$$

The parameters used in the model are summarized in Table 1. The continuous-time system must be approximated by a discrete-time system to carry out the computation of the output, making use of the discrete measurements – like the ones acquired from a scanner – as input to the system. The field of digital signal processing (DSP) offers many methods to accomplish this. It therefore helps to examine the model in the frequency domain by applying the continuous-time Fourier transform resulting in equation (2).

$$\begin{aligned} C_t(j\Omega) &= C_a(j\Omega) H(j\Omega) \\ H(j\Omega) &= \left[\frac{\left(\frac{K_{trans}}{1 - HCT} \right)}{K_{ep} + j\Omega} + V_b \right] e^{-\tau j\Omega} \\ H(j\Omega) &= [H_1(j\Omega) + H_2(j\Omega)] H_3(j\Omega) \end{aligned} \quad (2)$$

Examining the model in frequency allows the overall system to be broken down into the following 3 simpler parts: summation of constant gain V_b with a first-order system, $H_1(j\Omega)$ and an overall delay element $H_3(j\Omega)$. Note, the delay element is required because the measurement site is upstream to the input and it will take some amount of time for the contrast agent to arrive at the measurement site. Frequency domain analysis offers several discretization approaches—mainly finite impulse response (FIR) approximation and infinite impulse response (IIR).

The objective is to find parameters K_{trans}, K_{ep}, V_b, τ given the measurements $C_t[n]W_c[n]$ and $C_a[n]W_c[n]$. This is done using constrained nonlinear numerical optimization attempting to minimize the sum of square errors.

$$f(K_{trans}, K_{ep}, V_b, \tau) = \sum_{n=0}^{c/T} \left(\hat{C}_t[n]W_c[n] - C_t[n]W_c[n]\right)^2$$
$$\begin{aligned} 0 < \ & K_{trans} \leq 5 \\ 0 < \ & K_{ep} \leq 10 \\ 0 \leq \ & V_b \leq 1 \\ 0 \leq \ & \tau \leq c \end{aligned} \qquad (3)$$

Where $\hat{C}_t[n]$ represents samples of system output for a given set of parameters K_{trans}, K_{ep}, V_b, τ and a particular AIF$C_a[n]W_c[n]$. The summation limits reflect the fact that our measurements are cut off after $n = c/T$ samples. The optimization constraints were chosen to be within reasonable physical limits, and to aid certain optimization algorithms converge quicker.

Note that to compute $\hat{C}_t[n]$ the model (2) must be discretized. The discretization step introduces its own set of errors. In particular the choice of sampling rate and continuous-to-discrete mapping approach affect how well the discrete-time system resembles the continuous-time system at the range of frequencies of interest. The accuracy of fitted parameters depends greatly on the accuracy of the system approximating $\hat{C}_t[n]$.

Discrete Approximation Methods and Sampling Rates

There are 2 main methods evaluated in this paper to approximating the continuous-time system by a discrete system. The first method is the FIR using the window approach to filter design and the second is IIR using bilinear transformation (also known as Tustin's method). How well the discrete system approximates the continuous-time system depends largely on the sampling rate used during approximation (see online supplemental Appendix).

Although acquiring data at very high sampling rates is not clinically feasible, this section discusses the ideal signal processing case. Two factors affect the selection of appropriate sampling rate, both of which depend on the cutoff frequency - i.e., the point in the frequency domain where the signal is zero. Nyquist requires sampling rate to be at least 2x the cutoff frequency to avoid aliasing error (8). The second factor for selecting sampling rate is to ensure the discrete-time system matches the continuous system closely up to the cutoff frequency. Even if Nyquist rate criteria is satisfied, the discrete approximation may not match the continuous system up to the cutoff frequency and additional error may be introduced. In certain circumstances the acquired data should be up-sampled and processed at a higher rate to avoid introducing this additional error.

When the signals are not band limited and do not reach zero past any frequency, like in this case, a cutoff frequency is selected based on desired precision and computational feasibility. A low pass filter (LPF) is used prior to digitizing the signal to attenuate components past the cutoff frequency. The degree of attenuation in the stop band of the LPF depends on the noise floor, which is the background noise that is technically infeasible to get rid of in the system.

In the ideal simulation case where population average AIF is computed and then in turn used to generate signals, the noise floor is due to errors in floating point arithmetic. Studying the signals involved in the Tofts model, the cutoff frequency for the ideal case can be determined based on when the frequency components reach below the noise floor level (as if the low pass filter was applied). It was determined that to achieve precision on the order of single floating point arithmetic error, sampling rate of 3500 Hz is required (more detail can be found in the online supplemental Appendix).

Efficient Fractional Delay Approximation

As mentioned earlier, there is a delay between the time when the contrast agent is injected and when it arrives at the measurement site. This can be expressed as a continuous-time system $H_3(j\Omega)$. To account for this delay, the DCE analysis implementation could ask the user to visually evaluate the curves and supply the delay value when the tissue response curve begins to increase and optimize the other 3 kinetic parameters of the model; this approach would be tedious for a user to perform repeatedly for each voxel, error prone, as visual analysis could differ between users, and error prone if the user specifies the same delay value for a large physical area, which does not account for the fractions of seconds that it took for tracer to arrive at a further upstream site. Another approach to account for the delay could involve analyzing tissue response curves automatically based on the curve slope to determine the onset time, and then optimize the other 3 kinetic parameters (6). Heuristic search based on slope is susceptible to noise if there are noisy spikes before the true onset or if the onset occurs between samples. For this DCE analysis implementation, it was decided to numerically optimize all 4 kinetic model parameters, including the delay.

The discretization approaches, FIR and IIR, described in previous sections can deal with only delay by whole number of samples. For example if the system's sampling period is 1 s, only integer delay may be computed. This coarse approximation of delay can lead to poor fit in other parameters—K_{trans}, K_{ep}, V_b. The sampling rate can be increased to allow for a broader range of delay values—for example, 10 Hz would allow for any delay that is a multiple of 0.1 s—but at a proportional cost to memory requirements and processing time. This problem can be alleviated with the use of fractional delay approximation, which allows for estimation of the output signal for any floating point delay value (9). In our investigation the first-order Thiran filter considerably improved the results with negligible additional run-time cost. The delay in seconds can be implemented by the following 2 operations: Delay By Whole # of Samples

Table 2. Data Sets Analyzed

Name	Samples	Duration	Gaussian Noise
Data set 1	200 samples 1-second interval	200 seconds	None
Data set 2	9 samples 2-second interval	209 seconds	Added: $\mu = 0$
	19 samples 5-second interval		$\sigma = 6HU$
	9 samples 10-second interval		
DCE-CT Brain Scan	9 samples 2-second interval	209 seconds	Estimated: $\mu = 0$
	19 samples 5-second interval		$\sigma = 6HU$
	9 samples 10-second interval		

$N = \lfloor \tau / T \rceil$ followed by Fractional Delay $FD = \tau / T - \lfloor \tau / T \rceil$. The first-order filter is provided in equation (4)

$$H_{thiran}(z) = \frac{a_1 + z^{-1}}{1 + a_1 z^{-1}} \qquad a_1 = \frac{1 - FD}{1 + FD} \tag{4}$$

Testing Framework Design and Investigation Goals

The following were the investigation goals when designing the test framework:

1. Derive theoretical background for the ideal case to validate algorithm implementation and calibrate values for the basic numerical optimization algorithm parameters.
2. Investigate and demonstrate the effects of discretization method, sampling rates used during processing, noise, and fractional delay approximation filters on the resulting accuracy of the kinetic model parameters.
3. Investigate achievable accuracy of kinetic parameters extracted from clinical data set.

An experimentally derived functional form of population-average AIF (10) was sampled at 3500 Hz based on theoretical discussion in the section with the heading "Discrete Approximation Methods and Sampling Rates" in this paper. A uniformly distributed pseudorandom number generated was used to sample parameters K_{trans}, K_{ep}, V_b, τ from the minimization constraints range (8). The tissue curves were then calculated for each parameter set by a discrete-time system approximating the continuous model at 3500 Hz.

The ideal generated tissue curves proceed to a measurement stage where ideal high sampling rate signals are decimated and additional Gaussian white noise may be added. A summary of data sets analyzed and their canonical names used throughout the paper are summarized in Table 2.

In this setup, the ground truth parameters for data sets 1 and 2 are known. The generated signals at 3500 Hz represent the ideal case and it should be possible to recover the original parameters used to generate the signals to within tolerances of single floating point precision arithmetic. Running numerical optimization on the ideal signals was used to calibrate and configure the algorithms, as well as validate all additional custom code. The optimization algorithms evaluated in the simulation include: sequential quadratic programming (SQP) (11), downhill simplex (Nelder–Mead) (12), pattern search (PS) (13), simulated annealing (SA) (14), and differential evolution (DE) (15). Matlab (v2015b) optimization and global optimization toolbox's implementation of SQP, Nelder–Mead, PS, and SA were used. Price et al. implementation of DE was used for the experiments (15).

The algorithm parameters and values configured during calibration are described in Table 3. To overcome problems of local minima, SQP, Nelder–Mead, PS, and SA were initialized to

Table 3. Algorithm Parameters

Algorithm	# Start Points	Max Iterations	Exit Criteria	
			TolFun	TolX
SQP	32	1000	10^{-8}	10^{-8}
Nelder–Mead	32	1000	10^{-8}	10^{-8}
CUDA Nelder–Mead	32	1000	10^{-8}	10^{-8}
PS	32	1000	10^{-8}	NA
SA	32	1000	10^{-8}	NA
DE	64	1000	10^{-8}	NA
CUDA DE	512	1000	10^{-8}	NA

Table 4. Algorithm Calibration at 3500 Hz: Median of Percent Error and Timing

Algorithm	Overall %Error	Time (sec./voxel)
SQP	$8.97 \times 10^{-6} \pm 4.66 \times 10^{-7}$	1030±16
Nelder–Mead	$5.69 \times 10^{-8} \pm 2.32 \times 10^{-9}$	522 ± 23.7
CUDA Nelder–Mead (IIR)	$1.07 \times 10^{-7} \pm 1.27 \times 10^{-8}$	$(14.5 \pm 9.82) \times 10^{-3}$
DE	$3.27 \times 10^{-7} \pm 2.20 \times 10^{-8}$	1230 ± 12.3
CUDA DE (IIR)	$3.35 \times 10^{-7} \pm 2.59 \times 10^{-8}$	$(34.0 \pm 5.33) \times 10^{-3}$
PS	2.79 ± 1.04	13300 ± 284
SA	3.85 ± 1.23	2960 ± 32.5

quasi-random starting points generated using the Halton sequence (16). A quasi-random sequence was used to avoid the probability of generating tight clusters of starting points that could arise when using a distribution generated by a pseudo-random number generator. Each algorithm was configured to exit based on the maximum number of iterations, a minimum change in objective function (TolFun), and a minimum change in estimated parameter (TolX) to avoid infinite run-time. DE operates on a population of candidates that can conceptually be considered as the number of starting points. Furthermore, the DE objective function–based exit criteria was chosen such that the algorithm would exit when the difference between minimum and maximum values of the current objection function across the population was found to be below the TolFun threshold. All algorithm parameters were tweaked experimentally until the accuracy of the results were within the maximum accuracy allowable by a single floating point precision arithmetic or the results produced by the algorithm did not show any further improvement indicating numerical optimization algorithm limitations.

After calibration of the algorithm parameters (TolX, TolFun, etc.) and after having established an accuracy baseline, changes to objective function calculation in the form of adding fractional delay, changing discretization methods, and sampling rate were implemented. The validity of such code changes was verified by ensuring that at ideal processing rates, the accuracy matched the baseline accuracy. Then, data sets 1 and 2 were processed and the performance of each change was analyzed for its impact on accuracy and speed.

Analyzing the impact results, 2 algorithms were ported to CUDA to run on the GPU. In case of DE, the population size was increased to 512 compared to its CPU counterpart to take advantage of the multithreaded GPU architecture and have each optimization converge faster. Data sets 1 and 2 and an additional clinical DCE-CT brain scan were analyzed using this numerical optimization implementation under an institutionally approved REB protocol. The analysis was performed on CPU and GPU.

In terms of underlying hardware and timing analysis, the simulations were performed on several Xeon E5-2690 CPUs, and for comparison, on Tesla K40m GPU. A high-throughput computing cluster HTCondor was used; however, to narrow the analysis to only the algorithm performance, the overhead of data serialization, network transfer, and start-up time on remote nodes were discarded—only the main algorithm run-time was recorded.

In summary, earlier theoretical discussion led us to design for the ideal case under a single floating point precision. The algorithms were calibrated to perform within tolerances specified by the ideal case. With established confidence in correctness of implementation and calibration parameters, 2 artificial data sets were generated and run through the testing framework, while several other parameters were changed including the sampling rate and discretization method used on the Tofts model and the use of fractional delay approximation versus rounded delay for estimating the contrast arrival time at the site. Because, the second data set had the same sampling and noise profile of a scanned DCE-CT brain scan data set, when numerical optimization was carried out on the clinical data set, a conclusion on the accuracy of the extracted parameters could be determined.

RESULTS

Algorithm Calibration

The percent relative error for each parameter is defined as $\in = 100 \times |x_{true} - x_{approx}| \; / \; |x_{true}|$. The percent relative errors for each of the 4 parameters was combined into a single array of errors and the mean statistic along with 95% confidence interval was calculated and summarized in Table 4. Note that these calibrations are processed at very large sampling rates as discussed in the section with the heading "Discrete Approximation Methods and Sampling Rates" in this paper.

The SQP algorithm hits its optimization accuracy limit at percentage errors 1 and 2 orders of magnitude below DE and Nelder–Mead algorithms; decreasing tolerances and increasing sampling rates did not produce better results for SQP. The likely reason for this has to do with the fact that SQP is a gradient approach and the function is quite flat around the optimal point. This conclusion lead us to investigate nongradient-based approaches. From these approaches, Nelder–Mead and DE performed quite well. However, PS and SA could not be configured to achieve optimization values anywhere close to other algorithms; further modifications of algorithm parameters (such as increasing the number of starting points) produced marginally better results at a cost of much higher run-times. Because of these calibration results, long run-times and poor-accuracy PS and SA algorithm were discarded as viable numerical optimization candidates for this particular problem.

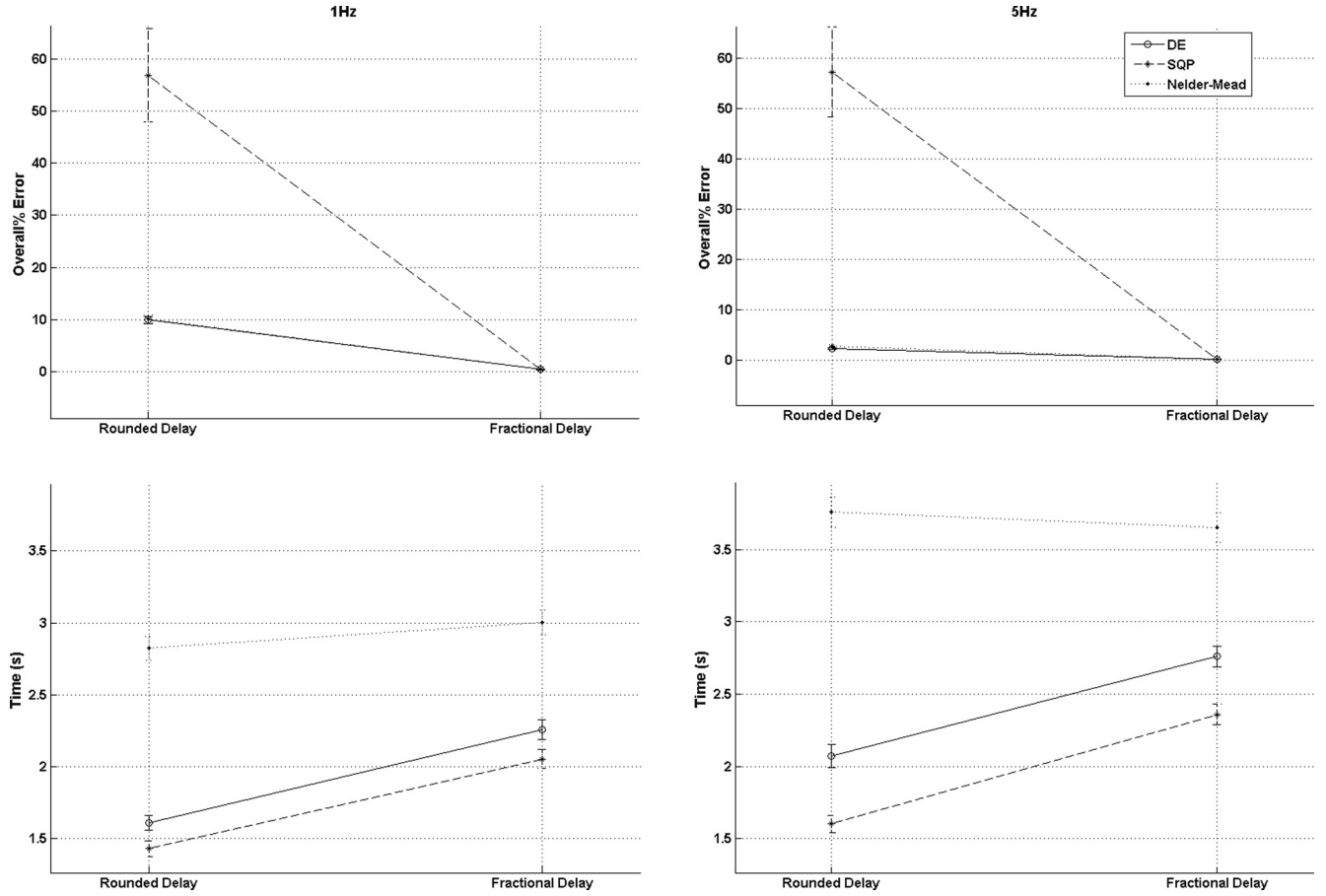

Figure 1. Data set 1. Impact of rounded delay vs fractional delay analysis processed at 1-Hz and 5-Hz infinite impulse response (IIR) on the mean overall %error and mean run-time per voxel.

Fractional Delay Analysis

Figure 1 shows the mean relative percent errors and the mean run-time sec./voxel with 95% confidence interval for results extracted from data set 1. IIR approximations at 1 Hz and 5 Hz were used. Using rounded delay approximation, SQP performs very poorly with the mean of the overall relative error at 56.9% regardless of the sampling rate used to process the data. One explanation for this is SQP exits criteria based on the objective function change is triggered because the gradient is constant for a range of delay values when rounding is used. Similar problems with rounded delay can be seen with DE and Nelder–Mead algorithm. With fractional delay approximation, instead of rounding, the error was reduced from 10% to 0.4% for DE and Nelder–Mead algorithms, and from 56.9% to 0.4% for the SQP algorithm.

An alternative to approximating the fractional delay is to use higher sampling such as 5 Hz. Somewhat surprisingly, SQP showed no improvement when using rounded delay compared to 1 Hz with the error still at 56.9%. The other numerical optimization algorithms did show a significant improvement where the overall error was 2.83%. However it should be noted that increasing the sampling rate by some factor increases the memory requirement by the same factor. Better accuracy can be achieved at 1 Hz with fractional delay approximation (0.4%) than at 5 Hz and using rounded delay (2.83%).

Figure 2 shows the fractional delay analysis run on data set 2, which has coarse, nonuniform sampling and additional $\mu = 0, \sigma = 6HU$ Gaussian noise added. Similar behavior can be observed for the SQP algorithm—it exits prematurely, causing very large errors (56.9%). Because of large amount of noise there (aliasing and artificial), there was no significant improvement in accuracy when using fractional delay approximation. It should be noted that in this case, the addition of the fractional delay approximation did not add significant amount of overall computation time.

In general, fractional delay approximation greatly improves accuracy of gradient-based numerical optimization algorithms such as SQP. When the noise profile of the data permits, it also improves accuracy significantly without having to process at higher sampling rates. Because of this, fractional delay approximation was added to all further analysis simulations and to the algorithms used to analyze clinical data.

Discrete Approximation and Sampling Impact Analysis

Figure 3 shows the means of relative percent errors across all parameters, as well as the mean logarithm of sec./voxel with

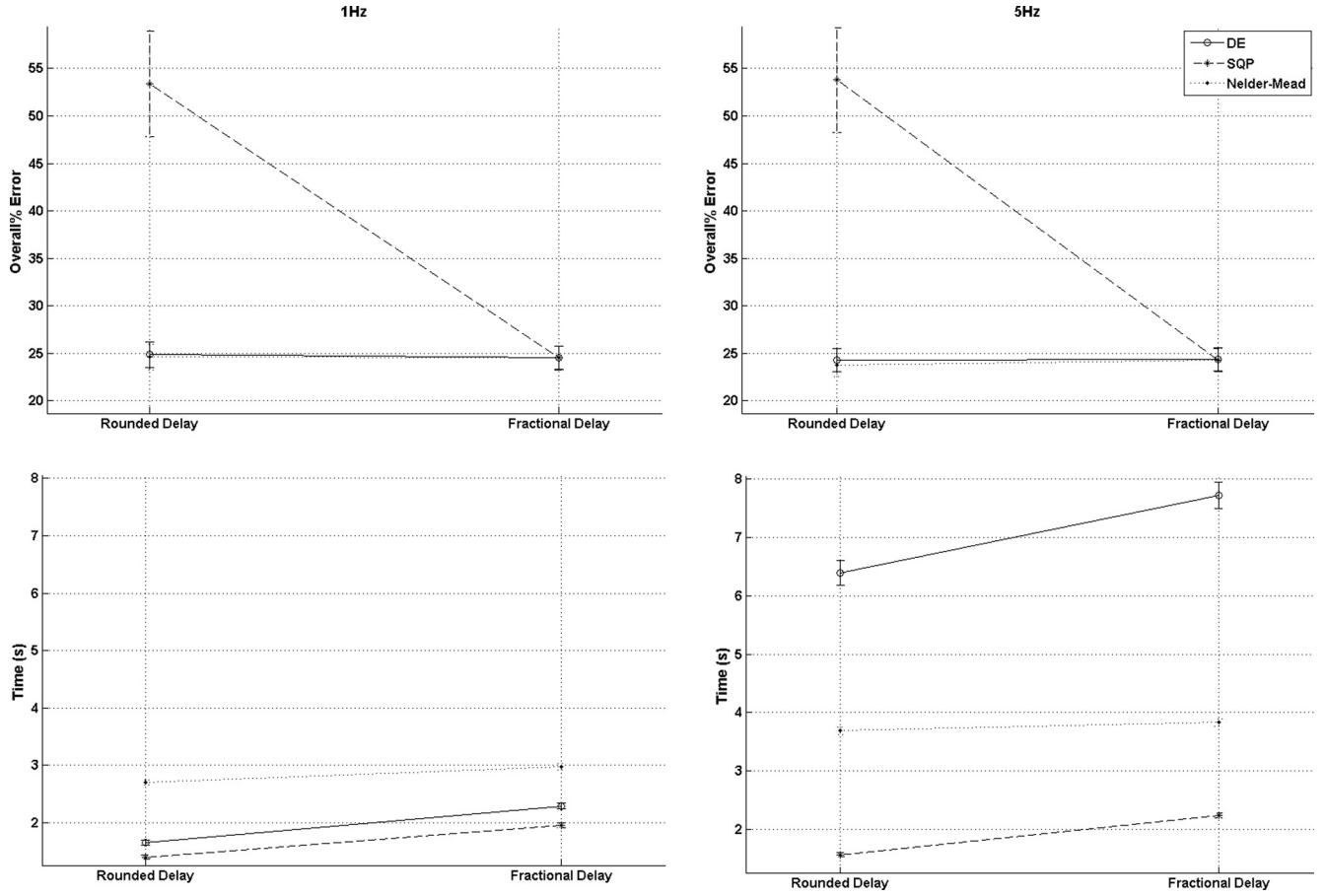

Figure 2. Data set 2. Impact of rounded delay vs fractional delay analysis processed at 1-Hz and 5-Hz IIR on the mean overall % error and mean run time per voxel.

95% confidence interval. The simulation compares 2 discrete approximation methods—FIR and IIR—and the effect of up-sampling data set 1 and using the more accurate discrete approximations that are a direct result of higher sampling rate. Fractional delay approximation was used during this analysis.

In terms of accuracy, the algorithms perform almost identically across sampling rates and discretization methods. Data set 1 was sampled at 1 Hz; the high-frequency information is lost forever regardless of how much the signals are up-sampled. However, if the signals were processed at 1 Hz, additional error would be introduced owing to the discrete-system poorly approximating the continuous-time system at this low rate. Figure 3 shows that accuracy can be increased by up-sampling the data and processing at higher rates. It is also evident that IIR approximation of the Tofts continuous-time system is more accurate than the FIR approximation at lower sampling rates, as the accuracy achieved by IIR approximation at 1 Hz is slightly better than the overall accuracy achieved by FIR approximation at 5 Hz. The mean of errors for each individual parameter when using IIR approximation at 1 Hz is {0.27%, 0.10%, 8.81%, 0.13%} for the parameters $\{K_{trans}, K_{ep}, V_b, \tau\}$, respectively. The overall mean error across all parameters is 2.33%. By switching to IIR approximation at 5 Hz, the overall mean of errors reduces to 0.40%, or individually, the error for each parameter becomes {0.46%, 0.16%, 0.84%, 0.12%}, showing large improvements for *Vb* parameter as a result of changing discretization method and increasing the sampling rate.

The run-time for the algorithm is shown as a log plot. For all sampling rates, IIR runs faster than FIR. The reason for this has largely to do with the fact that for this particular system, the IIR can be implemented in a single loop over the input data, so the complexity is $O(M)$, where M is the size of the signal. On the other hand, direct convolution requires 2 nested loops and has complexity $O(M^2)$. When signal size is large (such as when higher sampling rate is used), convolution implementation can be sped up by zero-padding the signals, computing the fast Fourier transform (FFT), multiplication of frequency bin values, and IFFT (17), in which case the complexity is $O(Nlog(N))$, where N is the size of padded signals. The implementation used during simulation uses the FFT approach, which handles larger signals much better than convolution. The algorithm complexity related to input size is evident in the timing plot, where FIR versions increase steadily as the sampling rate (and hence signal size) grows, whereas the IIR versions remain relatively flat.

The combination of better scalability as a result of algorithm complexity and the lower memory footprint requirement

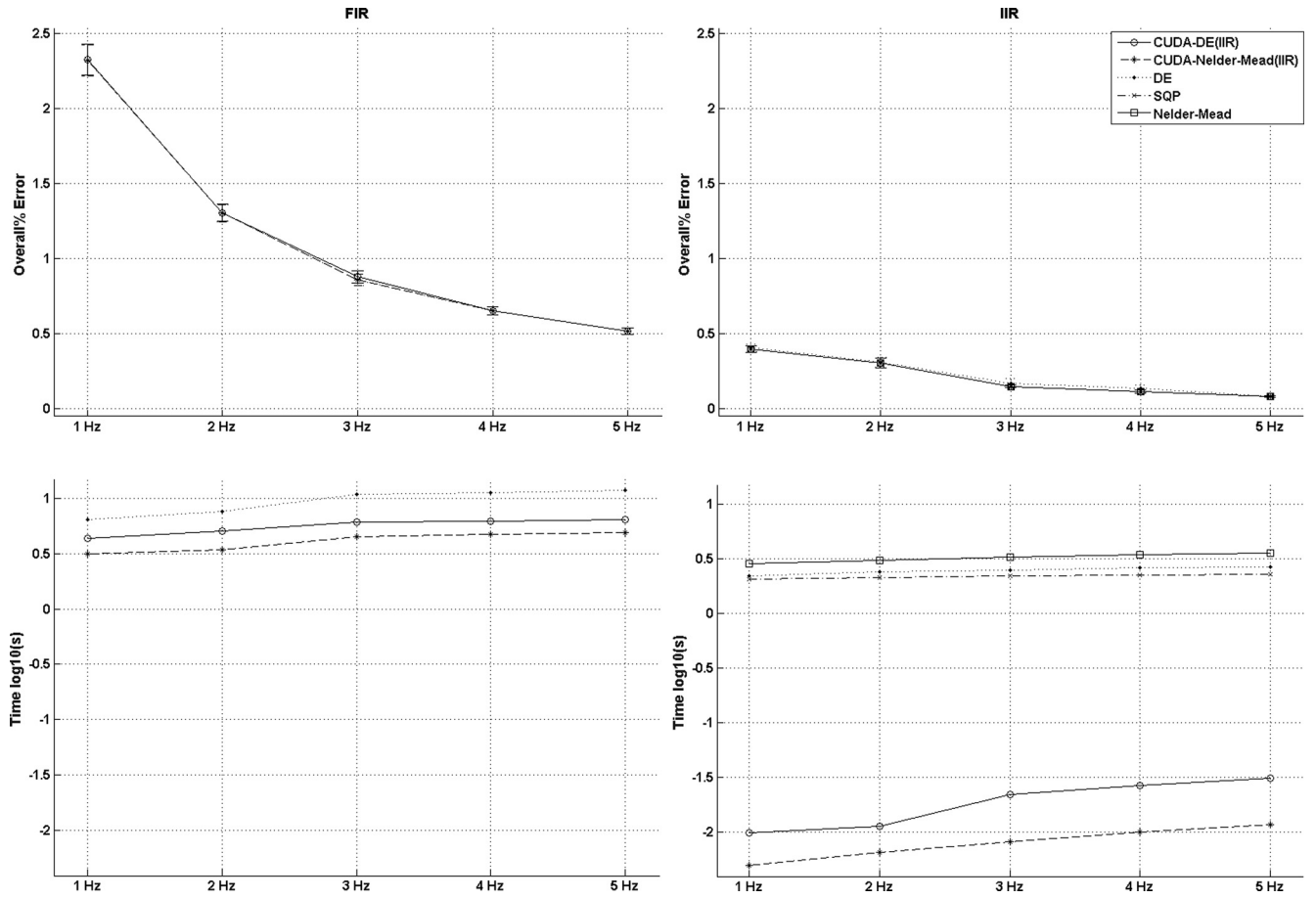

Figure 3. Overview of the impact of choice of sampling and discretization method on mean percentage overall error and mean run-time per voxel for Data set 1.

owing to better accuracy at lower sampling rates were the main reasons for using IIR approximation of the system in the CUDA implementation of DE and Nelder–Mead algorithms. The highly optimized CUDA implementation of the numerical optimization algorithms ran 2 orders of magnitude faster than their CPU counterparts.

Data set 2 was sampled nonuniformly, coarsely (average sampling rate 0.18) and had additional Gaussian noise ($\mu = 0, \sigma = 6HU$). Although the accuracy improvements from increased sampling and IIR approximation are very small, they are still evident. This analysis conveys the fact that data sets such as these need to be processed at only 1 Hz, as no further accuracy improvements can be gained by up-sampling to ensure the discrete-time system better approximates the continuous-time system. As a result of this analysis, the IIR approximation was chosen as the best discretization approach for this problem.

Figure 4 shows the results of the error analysis for data set 2 as a result of a changing the data sampling times. The resulting mean percentage error in parameter estimation was the smallest for the 1-s interval sampling interval and it increased with the increasing sampling rate. The clinical scan intervals varied depending on which part of the enhancement curve was being measured and the percentage errors therefore roughly correspond to the error values closest to the 3- and 5-s sampling intervals.

GPU Implementation and Clinical Data Analysis

Discrete approximation and sampling impact analysis showed that regardless of the optimization algorithm, IIR filter approximation produced more accurate results at lower sampling rates. In addition, fractional delay approximation allows for greater accuracy at lower sampling rates. Owing to excellent calibration accuracy, Nelder–Mead and DE, using IIR approximation and fractional delay filter, were chosen to be implemented in CUDA to run on the GPU. The calibration results from Table 4, along with identical accuracy compared to CPU counterparts (Figures 3 and 4), serve as verification that the algorithm implementation in CUDA is correct.

The best and fastest implementation (CUDA Nelder–Mead, with IIR filter and fractional delay approximation) was used to analyze a clinical DCE-CT brain scan. By analyzing CT scan areas that should contain a uniform CT number value, it was determined that the scanner may be adding as much as $\sigma = 6HU$ noise to the data. The noise was assumed to be Gaussian distributed (18) and the same population AIF was used as for the simulated curves. From earlier analysis on data set 2, which had

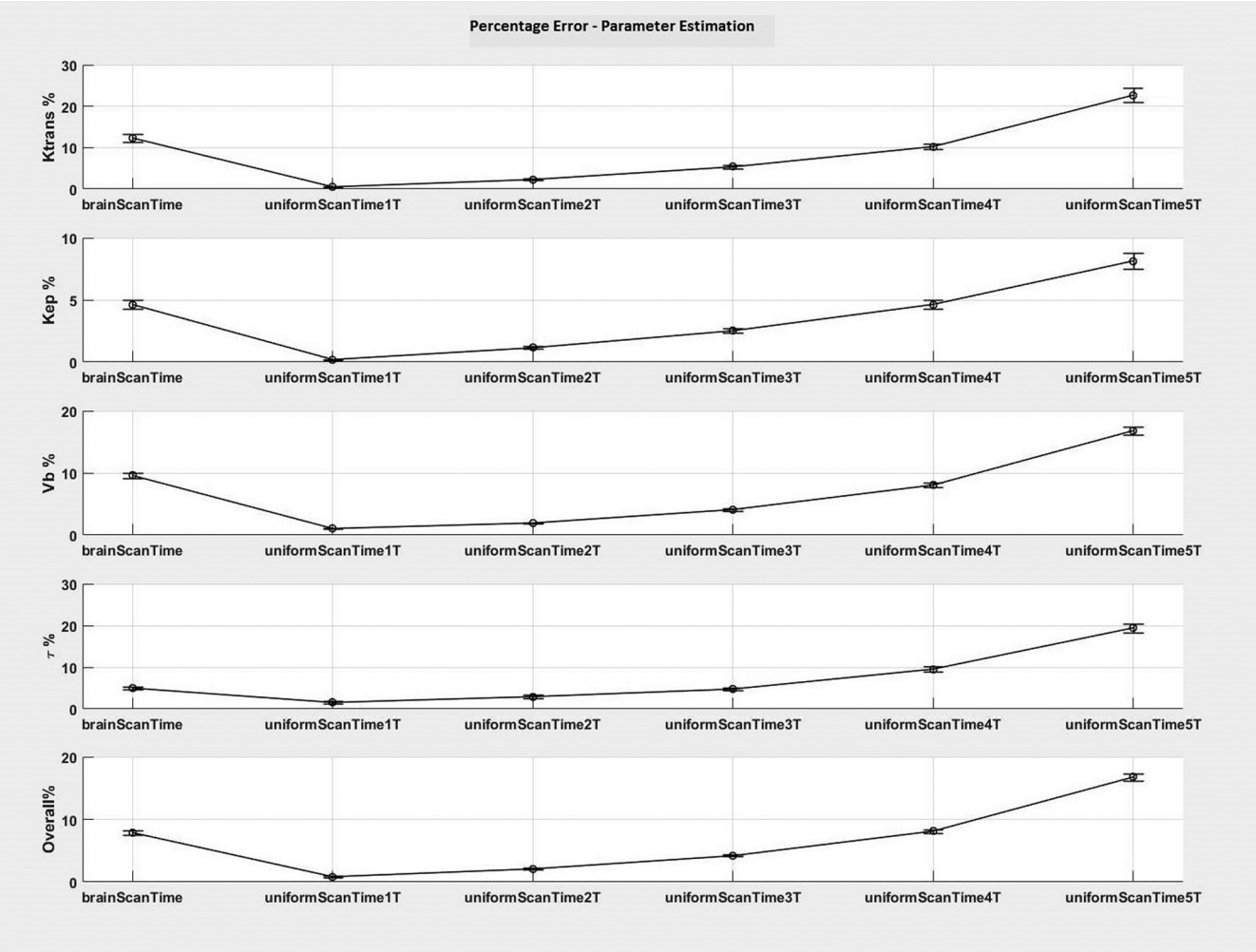

Figure 4. Data set 2, sampling analysis. Impact of data sampling on parameter estimation accuracy for (K_{trans}, K_{ep}, V_b, τ).

the same sampling and noise profiles as this CT data set, it can be concluded that the overall accuracy of parameters estimated from the CT data set is less than 10%.

Figure 5 shows a volume rendering of Vb parameter on the left, and the onset delay parameter rendering color coded such that red corresponds to earlier onset time and blue corresponds to later onset time.

Figures 3 and 4 show the GPU-based algorithm achieves speed improvements of 2 orders of magnitude compared with their CPU counterparts when run on generated data. Tables 5 and 6 show speed improvement when processing CT brain scan data. The first row is the baseline CPU implementation that uses FIR discretization of the Tofts model. The second row shows a modest speed increase because of changing the discretization to IIR. Finally the benefits of implementing the algorithm to run on a GPU are shown in the last row.

DISCUSSION AND CONCLUSIONS

Numerical optimization algorithms were carried out by designing for the ideal signal processing case at single floating point precision accuracy limits. Nelder–Mead, DE, and SQP produced good results under ideal conditions, achieving overall relative error 5.69×10^{-8}%, 3.27×10^{-8}%, and 8.97×10^{-6}%, respectively. SA and PS were found to be unsuitable for this problem because the lowest overall relative error that could be achieved was 3.85% and 2.79%, respectively.

The algorithms were designed and implemented to extract parameters from data sets with a wide range of sampling and noise profiles—ranging from the ideal and clinically infeasible data sets without noise to noisy and sparsely sampled CT brain data sets. To accomplish this, the thresholds for exit criteria were chosen to be of the order of 10^{-8}%. For very noisy data sets, this most likely creates a large amount of unnecessary processing that costs extra time; however, that is the trade-off to be able to achieve high accuracy for low-noise data sets as well. In cases of high-noise data sets, the numerical optimization exit is triggered when change in candidate parameter drops below threshold, rather than objective function target threshold. This is why for DE, the exit criteria were based on thresholding the difference between minimum/maximum objective function values across

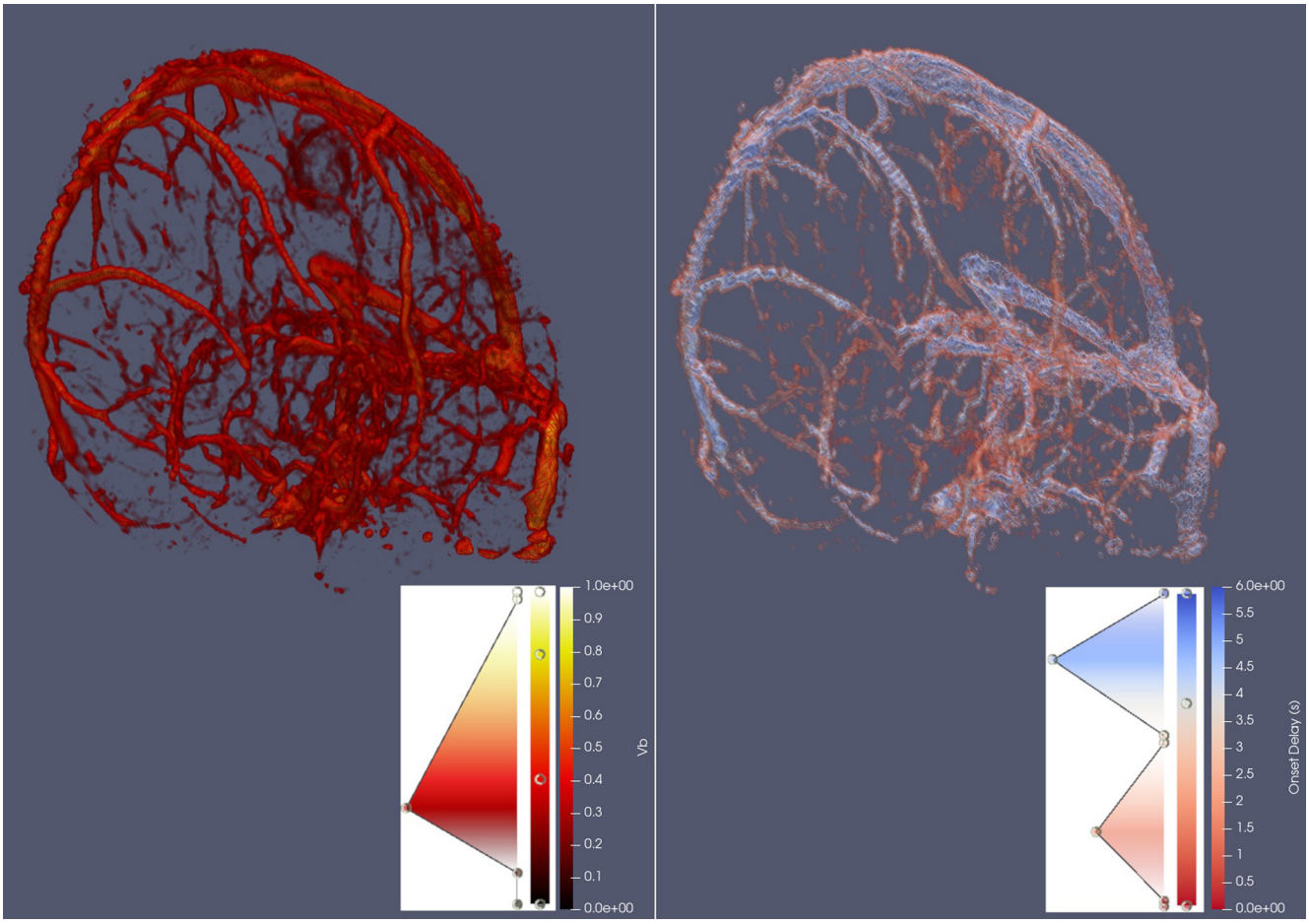

Figure 5. Volume rendering of Vb (left) and onset delay (right) parameters.

the population. Furthermore, numerical optimization algorithms that find local minima (compared to algorithms designed with global optimization in mind such as DE) were restarted many times at different initial starting points. Although the continuous objective function described in equation (3) may not have multiple minima, the discrete implementation of the objective function has many regions that would cause a numerical optimization algorithm to exit without reaching a point that would result in a better fit. For example, if rounded delay is used at 1-Hz sampling, the objective function is constant for all $\tau \in (0, 0.5)$, creating a saddle point which could cause numerical optimization to exit. This is especially evident in the gradient-based approach early termination summarized in Figures 1 and 2. Therefore as many as 32 starting points were used; using fewer starting points yielded poorer accuracy in the ideal optimization case. Having designed an algorithm that is capable of achieving best results in terms of accuracy for a very wide range of data, and a framework under which to conduct tests, it is possible to design a faster algorithm (by increasing thresholds of the exit criteria) that is able to achieve best results for the specific clinical data set.

Once numerical optimization algorithms were working to within designed tolerances of single floating point precision,

Table 5. Nelder–Mead Numerical Optimization CPU vs GPU Run-Time CT Brain Scan

Algorithm	Mean Time sec./Voxel	Relative Speed
CPU FIR 1 Hz	4.37	1.0
CPU IIR 1 Hz	3.05	1.4
CUDA IIR 1 Hz	0.0026	1680.8

Table 6. DE Numerical Optimization CPU vs GPU Run-Time CT Brain Scan

Algorithm	Mean Time sec./Voxel	Relative Speed
CPU FIR 1 Hz	2.51	1.0
CPU IIR 1 Hz	1.93	1.3
CUDA IIR 1 Hz	0.0068	369.1

experiments were conducted to vary other data processing steps and digital signal processing filters. It was shown that using fractional delay approximation filter stabilized gradient-based numerical optimization approaches and allowed the algorithm to produce accurate results instead of terminating early. Furthermore, fractional delay approximation allowed the discrete-time approximation for the Tofts model at lower sampling rates.

It was also shown that IIR discrete approximation of continuous-time Tofts model produces more accurate results at lower sampling rates. The recursive filter implementation has lower complexity compared to FIR discrete approximation, which requires convolution. This translates to lower memory footprint and faster processing times.

The clinical DCE-CT brain scan volume of interest contains just over 6 million voxels to analyze, after delineating and discarding areas outside the patient and bone. Combination of the 2 conclusions above led to an efficient port of the CPU-based algorithms into CUDA to run on the GPU. The framework can be used independent of image segmentation and run on every voxel or within a specific region of interest. The improvements in correlation between CT- and MRI-based measurements of tumor perfusion patients when a common analysis platform is used falls outside the scope of this article but is being reported on elsewhere (4).

To obtain entire brain perfusion maps required 4.3 hours (based on run-times in Table 5) on a single GPU; the same computation would take 179 days when processing on a single CPU (based on run-times reported in Table 6). If volume of interest is narrowed down further, for example, to only the tumor and surrounding tissue, which span 5 cc or just over 100,000 voxels, then kinetic model parameters can be computed in 4.3 min. Several orders of magnitude improvements such as these were also reported by Wang et al. (17) who achieved an even better 0.00025 s/voxel (compared to 0.0026 s/voxel) computation times using the block-FFT approach (FIR approximation of the Tofts model) on a less powerful GPU than Tesla K40. It should be noted that the implementation used for this paper used 32 starting points (effectively attempting to optimize each voxel 32 times to ensure global minimum) and stringent exit criteria. During CUDA code optimization attempts, it was found that the largest remaining barrier to even further speed optimization was noncoalesced memory access as a result of the delay parameter τ. In particular, on NVIDIA GPUs, the best speed can be achieved when the following holds: if a thread N reads memory location M, then thread N + 1 reads memory location M + 1 for all threads executing within a scheduled block. When implementing the delay which offsets the index of variables being read/written, coalesced memory access optimization does not apply, causing performance decrease.

A test framework such as this can further be used to determine the sampling rate required to process clinical data and gauge the magnitude of error that should be expected from the computed parameters, as well as calibrate numerical optimization algorithms to ensure best possible accuracy has been achieved.

Supplemental Materials

Supplemental Appendix: http://dx.doi.org/10.18383/j.tom.2018.00048.sup.01

ACKNOWLEDGMENTS

This work was supported by NSERC Discovery Grant #354701 and OICR operating grant #P.IT.020.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Jaffray DA, Chung C, Coolens C, Foltz W, Keller H, Menard C, Milosevic M, Publicover J, Yeung I. Quantitative imaging in radiation oncology: an emerging science and clinical service. Semin Radiat Oncol. 2015;25:292–304.
2. Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, Larsson HB, Lee TY, Mayr NA, Parker GJ, Port RE, Taylor J, Weisskoff RM. Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. J Magn Reson Imaging. 1999;10:223–232.
3. Driscoll B, Keller H, Jaffray D, Coolens C. Development of a dynamic quality assurance testing protocol for multisite clinical trial DCE-CT accreditation. Med Phys. 2013;40:081906.
4. Coolens C, Driscoll B, Foltz W, Svistoun I, Sinno N, Chung C. Unified platform for multimodal voxel-based analysis to evaluate tumour perfusion and diffusion characteristics before and after radiation treatment evaluated in metastatic brain cancer. Br J Radiol. 2018:20170461.
5. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, et al. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography. 2016;2:56–66.
6. Coolens C, Driscoll B, Chung C, Shek T, Gorjizadeh A, Menard C, Jaffray D. Automated voxel-based analysis of volumetric dynamic contrast-enhanced CT data improves measurement of serial changes in tumor vascular biomarkers. Int J Radiat Oncol Biol Phys. 2015;91:48–57.
7. Coolens C, Driscoll B, Foltz W, Pellow C, Menard C, Chung C. Comparison of voxel-wise tumor perfusion changes measured with dynamic contrast-enhanced (DCE) MRI and volumetric DCE CT in patients with metatstatic brain cancer treated with radiosurgery. Tomography. 2016;2:325–233.
8. Mani R, Oppenheim AV, Willsky AS, Nawab SH. Solutions manual, Signals & systems, Second edition. 462 p.
9. Laakso TI, Valimaki V, Karjalainen M, Laine UK. Splitting the unit delay [fir/all pass filters design]. Signal Processing Magazine IEEE. 1196;13:30–60.
10. Parker GJM, Roberts C, Macdonald A, Buonaccorsi GA, Cheung S, Buckley DL, et al. Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function for dynamic contrast-enhanced MRI. Magnetic Resonance in Medicine. 2006;56(5):993–1000.
11. Nocedal J, Wright SJ. Numerical Optimization. 2nd ed. New York: Springer; 2006.
12. Nelder JA, Mead R. A simplex method for function minimization. Comput J. 1965;7:308–313.
13. Audet C, Dennis JEJ. Analysis of generalized pattern searches. SIAM J Optim. 2002;13:889–903.
14. Press WH. Numerical Recipes: The Art of Scientific Computing. 3rd ed. Cambridge: Cambridge University Press; 2007.
15. Price K, Storn RM, Lampinen JA. Differential Evolution: A Practical Approach to Global Optimization: Verlag Berlin Heidelberg: Springer; 2006.
16. Halton JH. Radical-inverse quasi-random point sequence. Commun ACM. 1964; 7:701–702.
17. Wang H, Cao Y. GPU-accelerated voxelwise hepatic perfusion quantification. Phys Med Biol. 2012;57:5601–5616.
18. Coolens C, Breen S, Purdie TG, Owrangi A, Publicover J, Bartolac S, Jaffray DA. Implementation and characterization of a 320-slice volumetric CT scanner for simulation in radiation oncology. Med Phys. 2009;36:5120–5127.