Computational experimental design for quantitative MRI

by

Sean C. Epstein

a dissertation submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at

UNIVERSITY COLLEGE LONDON

April 2023

Supervised by:

Hui Gary Zhang

Timothy J.P. Bray

Margaret Hall-Craggs

I, Sean C. Epstein, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis presents contributions to the field of quantitative MRI (qMRI) computational experimental design (CED).

qMRI experiments are constructed from experimental 'building blocks' (e.g. acquisition protocol, model selection, parameter estimation) which, when combined, map tissue properties to quantitative biomarkers. Each of these blocks presents experimental choices: which acquisition protocol, which model, which parameter estimation method. Together, these choices form experimental designs. CED is the in-silico process by which such designs are tailored to suit specific imaging applications. This work addresses three limitations with current CED practices.

The first is that they are too narrow in their scope: they are unduly focused on acquisition protocol. qMRI is underpinned by model fitting, which relies on an appropriate choice of signal model and fitting method. This choice cannot be taken for granted: model fitting both depends on and influences the quality of the acquired data. This work argues that CED should not focus on acquisition protocol alone, but rather consider all experimental components in an end-to-end, holistic manner.

The second limitation relates to the experimental evaluation metrics currently used in CED. Experiments are assessed on their ability to generate close-to-groundtruth biomarker estimates, rather than on these estimates' ability to solve real-world tasks (e.g. tissue classification); there is a disconnect between evaluation and application. This work address this by proposing a CED method which assesses experiments on their task performance, and validates its assessments on two clinical datasets.

The final limitation relates to the parameter estimation methods available to CED. Existing methods are task-agnostic; they cannot be tailored to the needs of a specific qMRI experiment. This work takes advantage of machine learning techniques to, for the first time, make this possible: by changing training labels, parameter estimation performance is shown to be adjusted in a task-specific manner.

Impact statement

Magnetic resonance imaging (MRI) is a vital tool in modern clinical practice, offering non-invasive insight into a wide range of pathologies. Conventional MRI produces *qualitative* images, which are only sensitive to *differences* in tissue properties. In contrast, qMRI generates *quantitative* image 'maps' which not only record tissue differences but also *measure* tissue properties.

The information contained within such maps (not just *what* they measure, but also the accuracy and precision of those measurements) depends strongly on how a qMRI experiment is designed. There is no one-size-fits-all 'best' qMRI experiment; each use-case necessitates its own 'experimental design' process. To date, this process has focused on qMRI's *tissue-property-measurement* ability; qMRI experiments are optimised to generate high-accuracy measurements of underlying tissue properties.

This thesis argues that this approach is ill-suited to many qMRI applications. In clinical contexts, qMRI measurements are prized for their *sensitivity* to differences in tissue properties; their numerical value - how accurately they measure tissue properties - is of secondary importance. In light of this observation, this work makes the case for explicitly *task-driven* experimental design: adjusting qMRI experiments to maximise utility (e.g. clinical diagnosis) rather than numerical measurement.

This insight and analysis has impact both inside and outside academia. In academic research contexts, it represents a paradigm-shift in the *purpose* of qMRI experimental design. This should lead to the development of a new family of (potentially commercialisable) computational methods built around task-driven qMRI experimental design.

Outside academia, in clinical contexts, this thesis' theoretical contributions should lead to a rethink of what qMRI can, and should, offer. Making qMRI experiments task-driven has the greatest impact in non-research contexts, where qMRI techniques become (i) more useful and therefore (ii) more widely used.

The theoretical contributions presented throughout this thesis are supported by computational tools which actually *perform* task-driven experimental design. These tools have been made publicly available, and should enable both academics and clinicians to apply and develop task-driven experimental design practices suited to their own qMRI applications.

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **For a research manuscript that has already been published:**

   (a) **What is the title of the manuscript?** Task-driven assessment of experimental designs in diffusion MRI: a computational framework

   (b) **Please include a link to or doi for the work:** https://doi.org/10.1371/journal.pone.0258442

   (c) **Where was the work published?** PLOSONE

   (d) **Who published the work?** PLOS

   (e) **When was the work published?** October 8, 2021

   (f) **List the manuscript's authors in the order they appear on the publication:** Sean C. Epstein, Timothy J. P. Bray, Margaret A. Hall-Craggs, Hui Zhang

   (g) **Was the work peer reviewed?** Yes

   (h) **Have you retained the copyright?** Yes (CC BY license)

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi** https://doi.org/10.48550/arXiv.2103.08438

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   □ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** Sean C. Epstein: *Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review*

& editing; Timothy J P Bray: *Conceptualization, Project administration, Resources, Supervision, Writing – review & editing*; Margaret A. Hall-Craggs: *Project administration, Resources, Supervision, Writing – review & editing*; Hui Zhang: *Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing*

3. **In which chapter(s) of your thesis can this material be found?** Chapters 2 and 3.

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Sean C. Epstein
Date: 24/04/2023

Supervisor : Gary Hui Zhang
Date: 24/04/2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **For a research manuscript prepared for publication but that has not yet been published:**

   (a) **What is the current title of the manuscript?** Choice of training label matters: how to best use deep learning for quantitative MRI parameter estimation

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?** `https://arxiv.org/abs/2205.05587`

   (c) **Where is the work intended to be published?** Machine Learning for Biomedical Imaging (MELBA)

   (d) **List the manuscript's authors in the intended authorship order:** Sean C. Epstein, Timothy J. P. Bray, Margaret A. Hall-Craggs, Hui Zhang

   (e) **Stage of publication:** Submitted, under review.

2. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** Sean C. Epstein: *Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing*; Timothy J P Bray: *Conceptualization, Project administration, Resources, Supervision, Writing – review & editing*; Margaret A. Hall-Craggs: *Project administration, Resources, Supervision*; Hui Zhang: *Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing*

3. **In which chapter(s) of your thesis can this material be found?** Chapter 4.

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Sean C. Epstein

Date: 24/04/2023

Supervisor : Gary Hui Zhang

Date: 24/04/2023

Contents

Acronyms

ADC	Apparent diffusion coefficient 21, 29, 37–41, 46, 78, 94, 95
AIC	Akaike Information Criterion 22, 30
AUC	Area under the curve 34, 36, 39, 42, 45–48, 77–79
bcNLLS	Bound-constrained non-linear least squares 38–41, 78, 79, 110
BIC	Bayesian Information Criterion 22, 30
CED	Computational experimental design 18–20, 24, 25, 29, 33, 34, 36, 38, 41, 44, 45, 48, 49, 73–76, 78–86, 88
CNN	Convolutional neural network 87
CNR	Contrast-to-noise ratio 19, 20
CRLB	Cramér-Rao Lower Bound 20, 21, 24, 31, 91, 92
DL	Deep learning 51, 53–55, 58, 67, 70, 76–79, 81, 85, 97
dMRI	Diffusion MRI 20, 21, 29, 37, 38, 41, 48, 77, 93–95, 109
DNN	Deep neural network 51, 53, 61, 68, 72, 73, 84, 87, 97, 99–102
DTI	Diffusion tensor imaging 21
ELU	Exponential linear unit activation 58, 100
FIM	Fisher Information matrix 91, 92
IVIM	Intravoxel incoherent motion 37–41, 46, 47, 58–60, 64, 77, 78, 95, 109–111
ML	Machine learning 23–25, 74, 85
MLE	Maximum likelihood estimation 23, 27, 28, 53–63, 67–70, 73, 77, 78, 84, 106

MR	Magnetic resonance 18, 29, 86
MRI	Magnetic resonance imaging 17–20, 23, 27, 37, 38, 41, 42, 91, 105
MT	Magnetisation transfer 21
NLLS	Non-linear least squares 28
qMRI	Quantitative magnetic resonance imaging 18, 20–25, 27, 29, 30, 32–37, 41–46, 50–55, 58–60, 68, 70, 72–77, 79–88, 91, 93, 97, 102, 104, 105, 109
RL	Reinforcement learning 85–87
RMSE	Root mean squared error 21, 23, 30, 33, 60–62, 64, 65, 70–72, 106, 108
ROC	Receiver operating characteristic 34, 36, 37, 39, 42, 45–48, 77–80, 107
ROI	Region-of-interest 24, 35, 46, 106, 107
SIJ	Sacroiliac joint 37
sNLLS	Segmented non-linear least squares 38–41, 46, 79
SNR	Signal-to-noise ratio 20, 21, 23, 28, 35, 46, 58, 59, 61–65, 67–69, 71–73
SpA	Spondyloarthritis 37, 38, 45, 46, 77, 78
T1	Spin-lattice relaxation 20, 21
T2	Spin-spin relaxation 21
WLS	Weighted least-squares 38

Chapter 1

Computational experimental design for quantitative MRI

Contents

1.1 Introduction

Magnetic resonance imaging (MRI) is a vital tool in modern clinical practice, offering non-invasive insight which has revolutionised fields as diverse as oncology [1], cardiology [2] and neurology [3]. One of MRI's key strengths is its programmability and sensitivity to a wide range of biological phenomena. With the right software settings, MRI experiments can produce a near-limitless number of different image contrasts, enabling end-users to design imaging protocols that best-serve their specific experimental needs.

This flexibility presents a natural optimisation problem: selecting the MRI settings that produce the best, most useful, images. The simplest solution to this problem is empirical: multiple experiments are performed, each with different settings, and the resulting datasets are used to evaluate a *task of interest* (e.g. clinical diagnosis). The experimental settings which result in the greatest 'task performance' are selected as optimal.

Unfortunately, this process is time- and data-intensive; it is rarely feasible to evaluate more than a few different experimental settings in this way. Experimental designers may therefore wish to create shortlists of high-quality candidate experiments for subsequent empirical evaluation. This need has driven the development of computational experimental design (CED), which attempts to solve the experimental design optimisation problem in-silico, minimizing the need for real-world data.

This work focuses on the application of CED methods to quantitative magnetic resonance imaging (qMRI). qMRI is best understood by comparison with its 'conventional' qualitative counterpart. Conventional MRI describes experiments which output qualitative images: voxel intensities are dimensionless and image contrast reflects only *relative* information between nearby structures. These experiments use magnetic resonance (MR) scanners as cameras, acquiring images which show *differences* in tissue properties. In contrast, qMRI experiments acquire *stacks* of spatially-corresponding conventional MR images, and exploit the intensity relationships across these stacks to extract *quantitative* measures of tissue properties. This is achieved by introducing signal models which relate tissue-dependant *model parameters* to MR signals[1] and experimental settings. Settings are chosen, data is acquired, and parameters are estimated by fitting the signal model. In this way, scanners are converted from qualitative cameras into *measurement devices*, capable of quantifying tissue properties[2].

[1] For the purposes of qMRI CED, this Thesis treats MR scanners as imaging 'black boxes', controlled by acquisition settings θ_{acq}, which output fully-reconstructed noisy spatial maps. In this context, MR 'signals' refer to post-reconstruction voxel intensities (magnitudes). A detailed description of MR acquisition and reconstruction processes is readily available elsewhere [4].

[2] In clinical contexts, qMRI parameter estimates act as *biomarkers*, defined as 'characteristics that are measured as indicators of normal biological processes, pathogenic processes or responses to an exposure or intervention, including therapeutic interventions' [5].

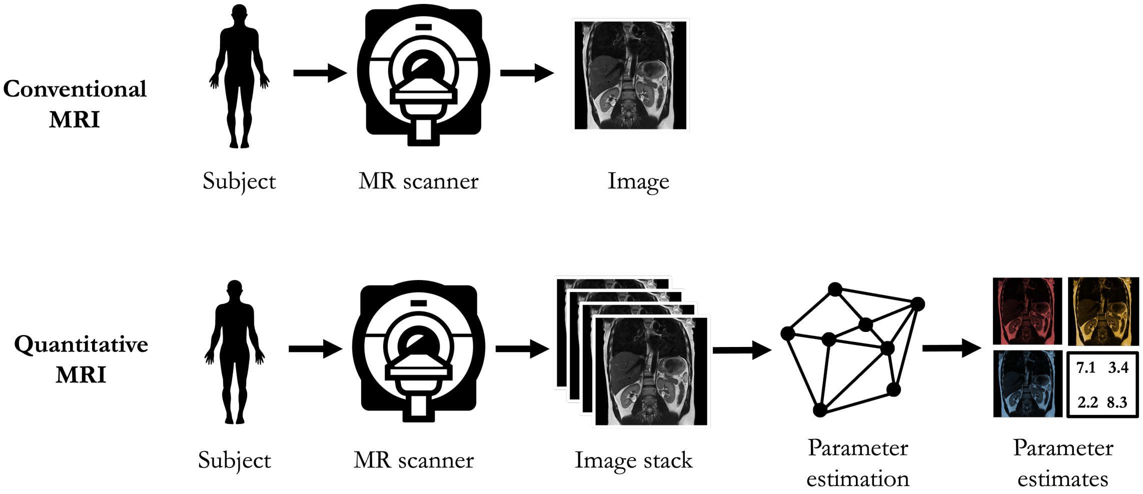

Figure 1-1: Overview of conventional and quantitative MRI

In conventional MRI, CED requires two components: a way to relate image acquisition settings and tissue properties to images (a *forward model*, or FM), and a way to assess the quality of these images (a *quality metric*, or Q). An exemplar forward model is the system of Bloch equations that describe MRI signal generation, and a corresponding quality metric could be the contrast-to-noise ratio (CNR) between two tissues of interest. In this example, computational experimental design might consist of (a) proposing acquisition settings θ_{acq}, (b) using the forward model FM to predict associated images I, (c) using the objective function OF to assess these images via quality metric Q, (d) adjusting the acquisition, and (e) iteratively repeating (a)-(d) until Q (image CNR) is maximised. This process is illustrated in Figure 1-2.

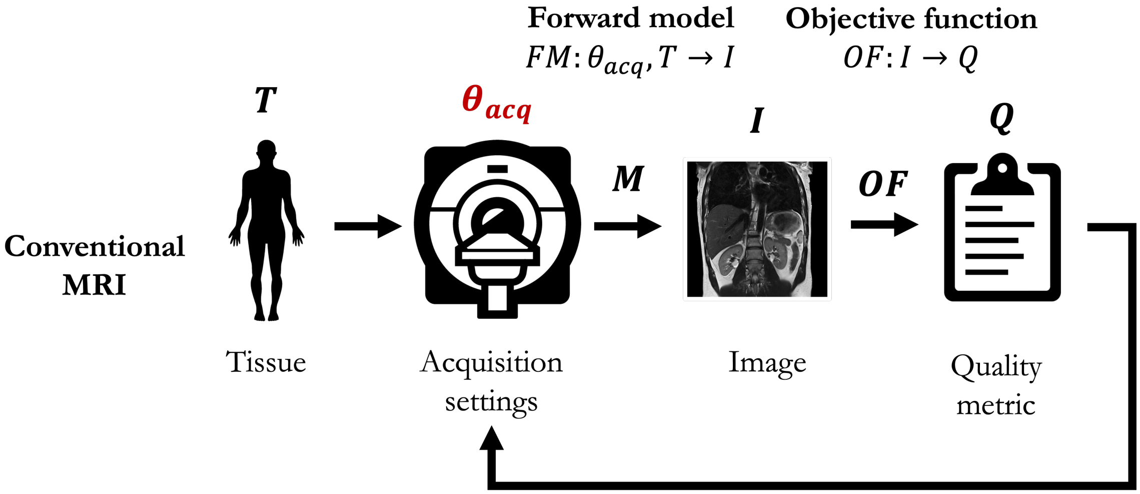

Figure 1-2: The computational experimental design algorithm in conventional MRI

In quantitative MRI, the CED process is more complicated. The qMRI forward model subsumes *three* experimental choices: not just which acquisition settings (θ_{acq}) to use, but also which signal model to fit (θ_{mod}) and which parameter estimation method to apply (θ_{est}). This thesis will describe the process by which these three choices have been made to date, critique it, and present actionable improvements to it. This work will, where necessary, base its analysis on diffusion MRI (dMRI), a widely used[3] form of qMRI, without loss of generality. dMRI quantifies the movement of water molecules over the typical length scales of cell microstructure[4].

1.2 Current practice

The three qMRI experimental design choices described above have historically been guided by the observation that, unlike in conventional MRI, qMRI experimental outputs (i.e. parameter estimates) relate to *physically meaningful* tissue properties. Therefore, qMRI CED need not rely on relative quality metrics like image CNR, but can rather target an *objective* metric: faithful reproduction of 'groundtruth' tissue properties. An extensive literature exists on this subject.

1.2.1 Choice of acquisition protocol (θ_{acq})

The first analytical framework for optimising qMRI acquisition schemes was proposed by Weiss et al. in the context of measuring spin-lattice relaxation (T1) [6]. This work derived an approximate analytic relationship between θ_{acq} and the variance of associated T1 estimates, valid at signal-to-noise ratios (SNRs) greater than 7, which was used to optimise θ_{acq}. Seven years later, a similar approach was applied by Wang et al. to more complex T1 multi-flip-angle experiments [7].

A significant advance in the field was made by Bain in 1990, again in the context of measuring T1 [8]. Bain proposed minimising biomarker variance by maximising the partial derivative of the acquired signal with respect to the biomarker of interest. In Bain's framework, the optimal θ_{acq} would be the one at which the acquired signal $S_{acquired}$ is most sensitive to changes in signal model parameters P. At these θ_{acq}, the ratio of $\Delta S_{acquired,\epsilon}$ (changes in signal due to noise ϵ) to $\Delta S_{acquired,P}$ (changes in signal due to differences in P) is minimised, and associated information-loss is reduced.

This approach was subsequently formalised within the framework of the Cramér-Rao Lower Bound (CRLB) by Jones et al. [9] in 1996. CRLB, which is described in detail in Appendix B, uses similar partial-derivative-maximising concepts to provide a θ_{est}-independent lower bound on biomarker variance. It is cheap to compute for even

[3] 604 Web of Science results for 'diffusion MRI' published in 2022, as of April 2023.

[4] See Appendix C for more details.

complex signal models, and has seen widespread use in a range of qMRI applications: T1 [10, 11, 12, 13], spin-spin relaxation (T2) [14, 15, 12, 13], magnetisation transfer (MT) [16, 17], and, since Brihuega-Moreno et al. [18]'s 2003 work, to dMRI. Following a comprehensive experimental design paper by Alexander et al. [19], it has become the most common data sampling optimisation tool within this field [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36]. For a signal model with n_P parameters, CRLB generates an $n \times n$ covariance matrix between each P (as per equation B.3). When it comes to 'minimizing' this matrix, a range of minimisation targets have been used: a weighted sum of its diagonal elements [16, 19, 13, 21, 23, 27, 37, 29, 31, 32, 35], its determinant [10, 26, 32, 33], specific diagonal elements [14, 15, 11, 36, 18, 22], an application-specific weighted-sum of specific elements [12, 17, 25, 30, 34], and a non-linear function of specific elements [20].

A range of non-CRLB numerical methods have also been applied to optimising θ_{acq}. Xing et al. [38] proposed using the SNR of qMRI parameter maps ('DNR') as an optimisation metric in the context of the apparent diffusion coefficient (ADC) dMRI model. They derived an analytic expression for this ratio, and decomposed it into terms relating to acquisition noise ϵ and qMRI 'sensitivity', which relates θ_{acq} to biomarker variance. Once derived, this sensitivity is straightforward to optimise for specific expected qMRI parameter values. Armitage et al. [39] extended this method to account for a range of parameter values, corresponding to multiple tissue types, by iterating over θ_{acq} until a target 'DNR' was achieved across all tissues.

Elsewhere, in contexts where analytic relations between θ_{acq} and biomarker error are not available, Monte Carlo optimisation methods have been applied. These methods iteratively minimise a loss function which depends either on θ_{acq} or parameter estimates $\overline{P}$.

θ_{acq}-based methods construct loss functions based on theoretically-derived (or intuited) notions of what constitutes 'good' qMRI acquisition protocols; these loss functions are minimised without reference to the associated qMRI biomarker estimates. Hasan et al. [40] contains a broad review of such methods in the context of diffusion tensor imaging (DTI) which, in contrast to the dMRI models discussed elsewhere in this thesis, encodes *directional* diffusion information in a diffusion *tensor*. In the context of DTI, an exemplar θ_{acq}-based Monte Carlo method is described by Jones et al. [41]: minimising the Coloumbic 'force' between the vector directions being acquired.

In contrast, $\overline{P}$-dependant Monte Carlo methods explicitly minimise errors associated with qMRI parameter estimation. This is achieved by (a) simulating synthetic qMRI data for which parameter ground-truths are known, (b) sampling this data using an acquisition protocol of interest, (c) fitting a signal model to this data, and (d) assessing the protocol by analysing the resulting parameter estimates. This protocol assessment has generally consisted of calculating bias and/or variance (often combined into root mean squared error (RMSE)) of the parameter estimates with respect to reference groundtruths [42, 43, 44, 45, 46].

1.2.2 Choice of signal model (θ_{mod})

The primary tension in qMRI model selection is the mismatch between (a) the complexity of the biological interactions that generate qMRI signals and (b) the relatively-low information content of the acquired data. This data can only support signal models which are gross simplifications of the underlying generative process. qMRI model selection has historically searched for the 'least simplistic' signal model, the one which best approximates the complex generative process. This search has usually relied on tools borrowed from the field of statistical model selection.

In most qMRI experiments, the number of acquired signals n_S is at most one or two orders of magnitude greater than the number of signal model parameters n_P, and assessing models on goodness-of-fit (e.g. fitting residuals) poses challenges; high dimensional signal models (large n_P) are able to improve goodness-of-fit by simply overfitting noise. For this reason, goodness-of-fit measures are commonly regularised by introducing terms which penalise n_P, as in the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) [47].

The AIC is defined as:

$$\text{AIC} = -2 \ln \mathcal{L}(\overline{P}, S_{acquired}) + 2n_P \tag{1.1}$$

where $\mathcal{L}$ is the likelihood of the best-fit parameters $\overline{P}$. A *smaller* AIC corresponds to a *better* model: the negative likelihood term $\mathcal{L}$ (encoding goodness-of-fit) is regularised by the positive model-complexity term $2n_P$. The BIC is structured similarly, but differs in its model-complexity term [47]:

$$\text{BIC} = -2 \ln \mathcal{L}(\overline{P}, S_{acquired}) + n_P \ln n_S \tag{1.2}$$

Both tools are data-driven: (i) qMRI data is acquired, (ii) maximum-likelihood parameter estimates are calculated for a range of signal models, and (iii) the models are ranked by either their AIC [48, 49, 50, 51, 52, 53] or BIC [54, 37, 49, 51] values.

An alternative, less common, model selection approach is *cross-validation*: selecting a model on its ability to predict unseen data. Ferizi et al. applied this technique, in combination with BIC, by fitting candidate signal models to 75% of acquired data and comparing the held-out data to the best-fit predictions [54]. In contrast, Rokem et al. took a test-retest approach: acquiring two scans of the same subject, and comparing a model's best-fit from one scan to the noisy data from the second [55].

1.2.3 Choice of parameter estimation method (θ_{est})

Parameter estimation is the process by which qMRI data is *mapped* to signal model parameter values which 'best represent' it. This involves (i) defining the mapping of interest (i.e. deciding what is meant by 'best represent') and (ii) computing this mapping for a particular noisy signal. As discussed above, the *mapping of interest* in qMRI is from data to groundtruth tissue properties which, outside of the machine learning (ML) contexts discussed in Chapter 4, is inaccessible. Instead, it is commonly approximated by a *proxy mapping* such as maximum likelihood estimation (MLE), which finds the parameter values associated with the highest probability of observing a particular qMRI signal. MLE can be straightforwardly performed using iterative algorithms available in many software packages; consequently, comparatively little attention has been paid to optimising θ_{est}.

Nonetheless, MLE's limitations[5] have led to the development of a range of competing parameter estimation methods [32, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68]. These estimators have historically been assessed on either (i) their approximation of the *mapping of interest* or (ii) their generation of the *proxy mapping* of choice.

Approximating the mapping of interest

This assessment paradigm compares qMRI parameter estimates to *reference groundtruth* values. These values, which are generally obtained in silico by synthesising testing data from known generative groundtruths[6], are used to calculate a range of quality metrics: RMSE [58, 60, 61], median error [62], mean absolute error [59, 63, 64, 65, 66], signal-space residuals [66], coefficient of variation [66], standard deviation of absolute error [63], and correlation with groundtruth [59].

Generating the proxy mapping

This form of assessment does not rely on parameter groundtruths. Rather, it evaluates how consistently an estimator is able to reproduce a chosen proxy mapping, such as MLE, *without assessing that mapping's relation to underlying tissue properties*. This assessment, which is generally applied to in vivo data, relies on a range of quality metrics: goodness-of-fit [60, 68, 69], intra-subject test-retest reliability [62,

[5] See Chapter 4 for a detailed discussion of limitations that have led to the development of machine learning parameter estimation methods. A notable MRI-specific limitation, not discussed elsewhere in this thesis, relates to the *noise model assumptions* underpinning MLE. Most model-fitting algorithms, when applied using default settings, assume the data being fit to has been corrupted by Gaussian noise (as described in Section 2.1). In reality, qMRI signals are subject to *Rician* noise [56]; this difference leads to biased MLE estimates at low SNRs [57].

[6] In some cases, as in Golkov et al. [58] and Liu et al. [59], groundtruth-like reference values are generated by calculating MLE estimates on super-sampled in vivo data.

64, 65], inter-subject region-of-interest (ROI) homogeneity [61, 62], intra-subject ROI homogeneity [62], intra-subject inter-ROI heterogeneity [60, 65], and consistency in inter-reader ROI means [61].

1.2.4 Summary

qMRI CED guides three experimental choices: acquisition protocol, signal model, and parameter estimation method. In making these choices, the existing orthodoxy can be summarised by (a) the quality metrics used, (b) the relative importance ascribed to each choice, and (c) the order in which they are made.

Quality metrics are based on viewing qMRI experiments as tissue-measurement tasks: experimental outputs (parameter estimates) are assessed on their similarity to tissue property 'groundtruths'; experiments are designed to minimise associated bias and/or variance.

More attention is paid to optimising acquisition protocol than signal model or parameter estimation. Low-bias signal model selection is trivialised by data-driven information metrics, and, up until the recent interest in ML techniques, the limited choices available in parameter estimators left little scope for optimisation.

The three qMRI experimental design choices are made *sequentially*: a model is chosen, an estimation method is selected, and only then is an acquisition protocol optimised. In the case of CRLB, the most popular protocol-optimisation technique, the interaction between acquisition protocol and parameter estimation is ignored.

1.3 Limitations of current practice

These approaches have two broad limitations.

The first relates to the *quality metrics* described above. In research settings, qMRI is used to gain new insight into tissue properties, and in this context it is appropriate to assess experiments on their ability to generate close-to-groundtruth parameter estimates. The issue arises in clinical settings, where qMRI is useful not for its measurement properties but for its high sensitivity and specificity in distinguishing tissues in different states of pathology. In these contexts, existing quality metrics manifest a disconnect between evaluation and application; qMRI CED should instead employ quality metrics which directly target task performance.

The second limitation relates to the *overall CED process*, as it exists today: the three experimental choices are made sequentially, often in isolation, without considering the complex non-linear interactions between them. The true value of a qMRI experiment, its task performance, depends on an appropriate *combination* of acquisition, model, and estimation. It is not meaningful to optimise any one of these choices without simultaneously considering the interaction with the other two.

1.4 Contributions

In light of these limitations, this thesis presents four contributions to the field of qMRI CED:

1. A conceptual re-framing of the purpose of CED, away from *tissue measurement fidelity* to a broader notion of *task performance maximisation.* This theoretical contribution is underpinned by a computational implementation, which shows how a qMRI experiment can be assessed on its ability to perform a clinical task (pathology subtyping).
2. A clinical validation of the previously described computational implementation, showing that by modelling the interactions between the three qMRI CED experimental choices, clinical task performance can be accurately and reliably predicted in silico.
3. A novel ML-based parameter estimation method, which, by complementing existing estimation methods, provides an additional tool for qMRI CED to maximise task performance.
4. An implementation of a 'task performance comparison' between traditional parameter estimation and a range of recently-introduced ML-based methods. This not only elucidates the trade-offs between these competing approaches, but also provides a blueprint for how to make experimental design choices in any qMRI CED context.

These contributions are described, sequentially, in the following four chapters.

Chapter 2

Re-imagining the problem: holistic & task-driven

Contents

This chapter is adapted from:

- **Epstein SC**, Bray TJP, Hall-Craggs MA, Zhang H, *Task-driven assessment of experimental designs in diffusion MRI: a computational framework*, 2021, PLOS ONE 16(10): e0258442

- **Epstein SC**, Bray TJP, Hall-Craggs MA, Zhang H, *Variability from complexity: assessing IVIM acquisition schemes through parameter estimation uncertainty*, 2020, ISMRM Annual Meeting 2020

- **Epstein SC**, Bray TJP, Hall-Craggs M and Zhang H, *Towards a computational framework for task-driven experimental design*, 2021, ISMRM Annual Meeting 2021

2.1 Theoretical background

Magnetic resonance imaging, in all its forms, produces spatial maps. In conventional MRI, these maps are qualitative: the numerical value (*intensity*) corresponding to a spatial location (*voxel*) is only meaningful in the context of the other voxels in the same map. Contrast is the only store of information in such maps ('images'), and is therefore the only target for optimisation during experimental design.

In contrast, the maps produced by qMRI experiments are quantitative, and information is stored in two forms: not just *relative contrast* but also *absolute value.* Voxel intensities have units and encode physically meaningful tissue properties.

Two key assumptions underpin the generation of these dual-information maps. The first is that a deterministic relationship $\mathcal{M}$ exists between tissue properties T and acquisition settings θ_{acq}, such that (with a suitably chosen θ_{acq}) different tissues will generate different signals:

$$S_{deterministic} = \mathcal{M}(\theta_{acq}, T) \tag{2.1}$$

The second assumption is that $\mathcal{M}$ can be approximated by a simplified signal model M, parameterised by P:

$$S_{deterministic} = \mathcal{M}(\theta_{acq}, T) \approx M(\theta_{acq}, P) \tag{2.2}$$

Now consider the acquired qMRI signal $S_{acquired}$:

$$S_{acquired} = \mathcal{M}(\theta_{acq}, T) + \epsilon \equiv S_{deterministic} + \epsilon \tag{2.3}$$

where ϵ is a noise instantiation that corrupts $S_{deterministic}$. Under the above assumptions, fitting M to $S_{acquired}$ generates parameter estimates $\overline{P}$ which contain *information* about T. This fitting is traditionally performed via MLE, which maximises a likelihood function $\mathcal{L}$ describing the probability $\mathcal{P}$ of observing $S_{acquired}$ given some parameters P:

$$\mathcal{L}(S_{acquired}, \theta_{acq}|P) = \prod_{i=1}^{n_S} \mathcal{P}(S_{acquired,i}, \theta_{acq}|P) \tag{2.4}$$

where i indexes the n_S signals contained within $S_{acquired}$, and it is assumed each $S_{acquired,i}$ is drawn independently from the same noise distribution. MLE identifies the P which maximises $\mathcal{L}$:

$$\overline{P} = \arg\max_{P} \mathcal{L}(S_{acquired}|P) \tag{2.5}$$

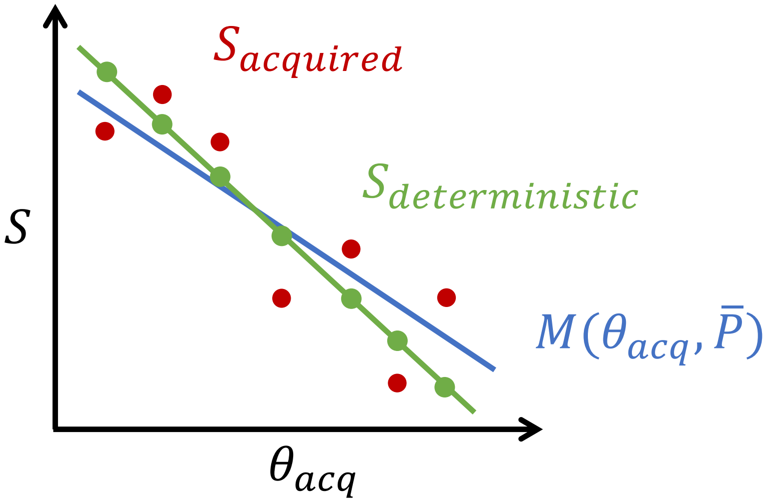

Figure 2-1: The parameter estimation problem, visualised for a simple linear model.

Under a Gaussian noise model[1], this maximization simplifies to the commonly-used non-linear least squares (NLLS):

$$\overline{P} = \arg\min_{P} \sum_{i=1}^{n_S} \|M(\theta_{acq}, P)_i - S_{acquired,i}\|^2 \tag{2.6}$$

More generally, the model fitting process (*parameter estimation*) can be thought of as a minimization of a difference metric $\mathcal{D}$ between $\overline{P}$ and a desired 'target' P_{target}:

$$\arg\min_{\overline{P}} \mathcal{D}(\overline{P}; P_{target}) \tag{2.7}$$

[1] As noted in Chapter 1 footnote 5, qMRI data is generally Rician, rather than Gaussian, distributed; these distributions converge at high SNRs, and a Gaussian approximation is commonly employed during parameter estimation.

where

$$P_{target} = \Phi(T; \theta_{est}, \theta_{mod}, \theta_{acq}) \tag{2.8}$$

and Φ is a lossy transform which depends on experimental settings θ. In other words, by minimizing the difference between $\overline{P}$ and P_{target}, parameter estimates are generated which encode information about tissue properties T. Performing this fitting process at every spatial location within an MR image generates a spatially-encoded *parameter map*, which stores T-derived information at every location[2].

This information takes two forms: *relative contrast*, which relates to differences in T; and *absolute value*, which relates to numerical values of T. These two forms of information are usually related but are, in general, distinct: an 'ADC parameter map'[3] may give *biased* diffusivity measurements whilst still reliably distinguishing tissues with *different* diffusivities.

Three experimental settings affect the information content of qMRI parameter maps: signal model (θ_{mod}), parameter estimation method (θ_{est}), and acquisition protocol (θ_{acq}). In general terms, the aim of qMRI experimental design is to adjust these experimental settings to maximise the information encoded by the resulting experiment.

This leads to two fundamental questions which underpin qMRI CED: (i) which form of information should be prioritised? and (ii) how should this be done?

2.2 Status quo: sequential & measurement-driven

To date, qMRI CED has focused exclusively on encoding the second, *absolute*, form of information. In other words, $P_{target} = T$; acquisition settings, signal models, and estimation methods have been chosen so as to produce $\overline{P}$ which are *accurate measurements* of underlying tissue properties T, rather than highly contrasting ones.

In pursuit of this tissue-measurement goal, qMRI CED has followed a *sequential* 3-step process: tissue characterisation, then model fitting selection, and finally acquisition optimisation.

[2] See Appendix F for a concrete example which grounds the notation introduced above.

[3] ADC refers to apparent diffusivity coefficient, a highly simplified approximation of the mean diffusivity of H protons within a dMRI voxel; see Appendix C for more details.

2.2.1 Tissue characterisation

For $\overline{P}$ to accurately measure T, it is generally assumed[4] that M should well-approximate $\mathcal{M}$. To this end, a range of M are compared, and the 'best-matching' one is selected.

High quality experimental data underpins this comparison. Such data, which may be oversampled with respect to routine qMRI acquisitions, is acquired in such a way as to contain as much information on $\mathcal{M}$ as possible. A number of candidate M are fit to it, and the quality of fit is assessed using an information-theoretic quality metric (Q_{mod}) such as AIC or BIC. The model which best fits the data is selected.

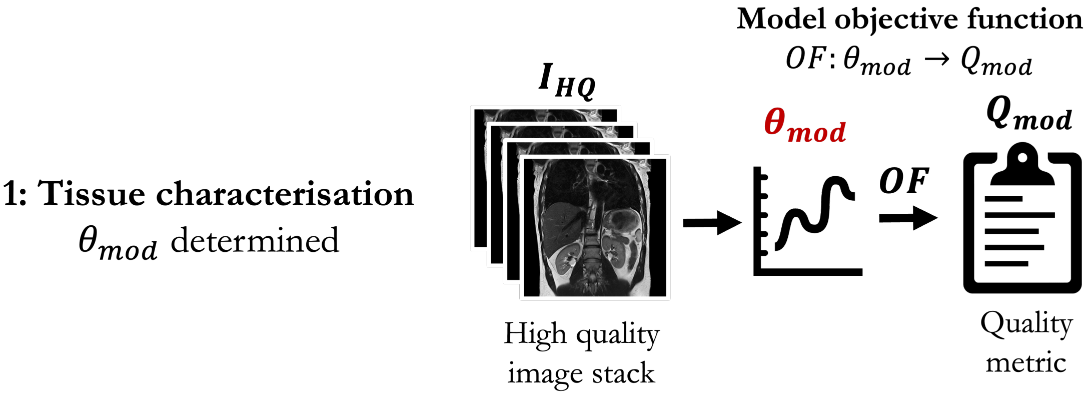

Figure 2-2: Tissue categorisation: choosing a signal model.

2.2.2 Model fitting selection

Having chosen a signal model, an accompanying parameter estimation method is selected. A range of candidate methods are proposed and compared. The test dataset underpinning this comparison generally contains reference 'groundtruth' P values ($P_{groundtruth}$) against which each method's $\overline{P}$ can be assessed. These groundtruths are often obtained by generating synthetic test data from known P. The method which minimises a difference-based quality metric (Q_{est}), such as RMSE, between $\overline{P}$ and $P_{groundtruth}$ is selected.

[4] This is not, in general, true. A highly-simplified model, which 'averages out' multiple confounding tissue property interactions, may accurately predict *some* elements of T.

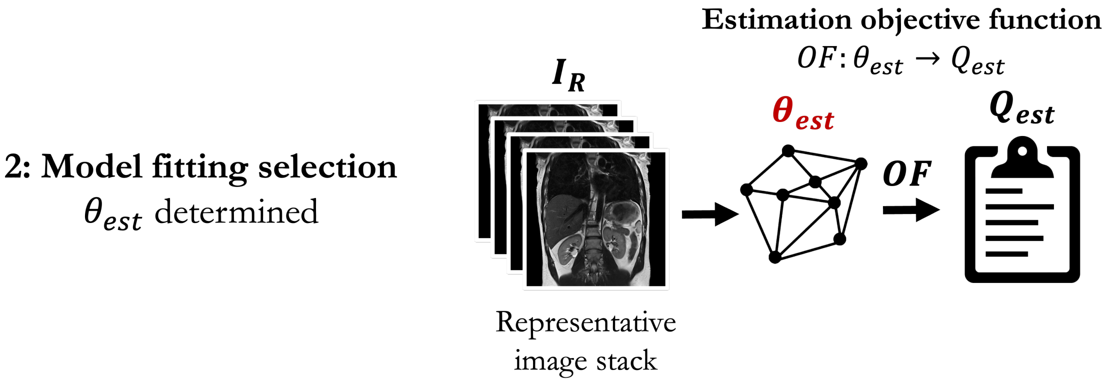
Estimation objective function
$OF: \theta_{est} \rightarrow Q_{est}$
I_R
θ_{est}
Q_{est}
2: Model fitting selection
θ_{est} determined
OF
Representative image stack

Figure 2-3: Model fitting selection: choosing a parameter estimation method.

2.2.3 Acquisition optimisation

Finally, acquisition settings are determined. In some cases, θ_{acq} are selected (independently of θ_{est}) by optimising a CRLB-derived metric. In others, candidate θ_{acq} are explicitly proposed and compared; each θ_{acq} is used to generate synthetic test data, and the $\overline{P}$ corresponding to the previously-selected θ_{mod} and θ_{est} are assessed against $P_{groundtruth}$, using quality metrics (Q_{est}) similar to those employed in model fitting selection. The acquisition settings that optimise this metric are selected.

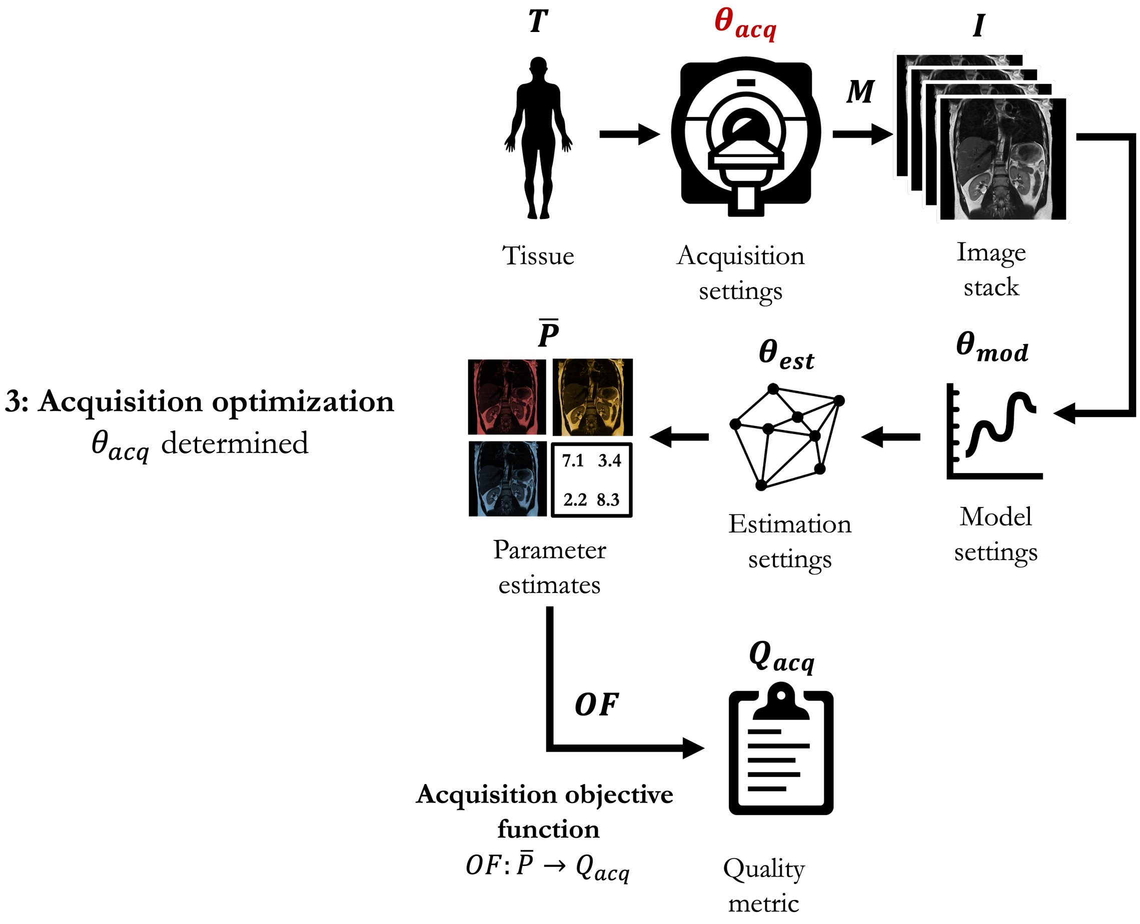

Figure 2-4: Acquisition optimisation: choosing acquisition settings.

2.3 A new approach: holistic & task-driven

The experimental design process described above is characterised by (a) its focus on encoding *absolute* (rather than *relative*) information content and (b) the linear, sequential nature by which θ_{mod}, θ_{est}, and θ_{acq} are selected.

However, both of these properties are ill-suited to experimental contexts involving tasks, such as tissue classification, where experimental outcome is measured in terms of *effect size* rather than measurement accuracy. In such settings, model parameters P should be treated as task-specific *biomarkers*, rather than tissue-specific *measurements*.

Implementing this change would correspond to a simple adjustment of P_{target} (Equation 2.8), the parameter estimates an ideal qMRI experiment would generate.

The updated P_{target} contains more relative information about T, *even if this comes at the cost of reduced absolute information.* In the example of tissue classification, Φ becomes the transform that maximises *differences* between tissues of interest, rather than maximising tissue property measurement accuracy.

This should not be a groundbreaking argument, as gold-standard non-computational qMRI techniques are similarly task-driven: experimental designs are assessed on their ability to correctly classify subjects with known diagnoses [67, 70]. Despite this, no CED task-driven framework exists.

Indeed, current CED approaches are fundamentally incompatible with task-specific Φ. Not only must quality metrics change (e.g. from RMSE to task performance), but the way in which these metrics are assessed must fundamentally change too. Sequential assessment only makes sense in a world of low-bias P_{target}; if there is no universally optimal M, θ_{mod} cannot be evaluated independently of θ_{est} or θ_{acq}. It is the *interaction* between these three experimental choices that must be assessed on its associated task performance.

In light of this, this thesis present a new qMRI CED theoretical framework, shown in Figure 2-5.

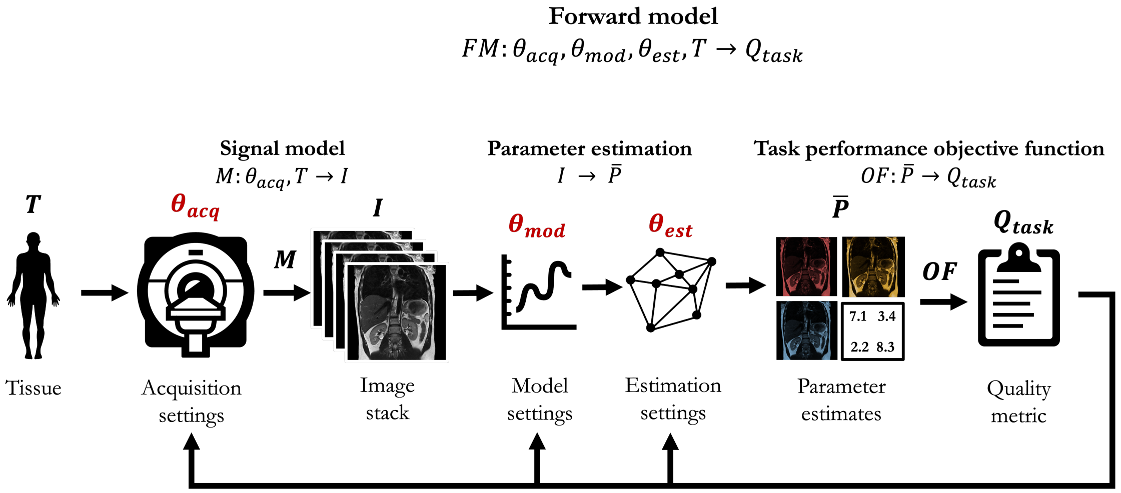

Figure 2-5: A novel qMRI CED paradigm.

In this framework, CED optimises the forward model FM which maps tissue properties and experimental design choices to task performance:

$$FM : T, \theta_{acq}, \theta_{mod}, \theta_{est} \rightarrow Q_{task} \tag{2.9}$$

where Q_{task} is a task-specific task performance quality metric. FM subsumes all experimental design choices, and explicitly assesses the *interaction* between these choices. This CED framework is flexible and backwards-compatible: if the task is defined as ‘tissue measurement’, P_{target} can be simply reverted to T.

2.4 Proof of concept implementation

In what follows, a proof of concept implementation (the ‘pipeline’) of this theoretical framework is proposed, implemented, and evaluated.

2.4.1 Proposal

Given a quantitative task and set of experimental design choices, the pipeline predicts task performance using quantitative summary metrics. For classification tasks, these metrics are typically receiver operating characteristic (ROC) curves and their associated area under the curve (AUC). The pipeline accommodates other metrics that may be more appropriate for a given application.

The pipeline mimics, in-silico, empirical task-driven assessment: gathering real-world data and measuring sensitivity and specificity. Its structure is shown in Figure 2-6; it takes three inputs: a quantitative task (I1), a characterisation of relevant tissue(s) (I2), and a candidate experimental design (I3). The pipeline combines these inputs to predict associated task performance. It simulates complete qMRI experiments: noisy qMRI data is synthesised (P1), qMRI model parameters are estimated (P2), and task performance is evaluated (P3).

Inputs

I1: A quantitative parameter-driven task, characterised by an operational definition and quantitative performance metric.
Example: Classification of tissue as either healthy or diseased, based on qMRI parameter estimates: a parameter is chosen, a threshold is set, and a tissue is diagnosed based on its estimated model parameter(s). Task performance is measured by the AUC of an ROC curve computed across a patient population (by sweeping the parameter threshold across all values).

I2: Characterisation of the tissues involved in the quantitative task (I1).
Example: The ‘ground-truth’ qMRI model that has been deemed to most-faithfully represent the underlying tissue; associated empirical parameter values (e.g. mean & standard deviation) of each tissue type.

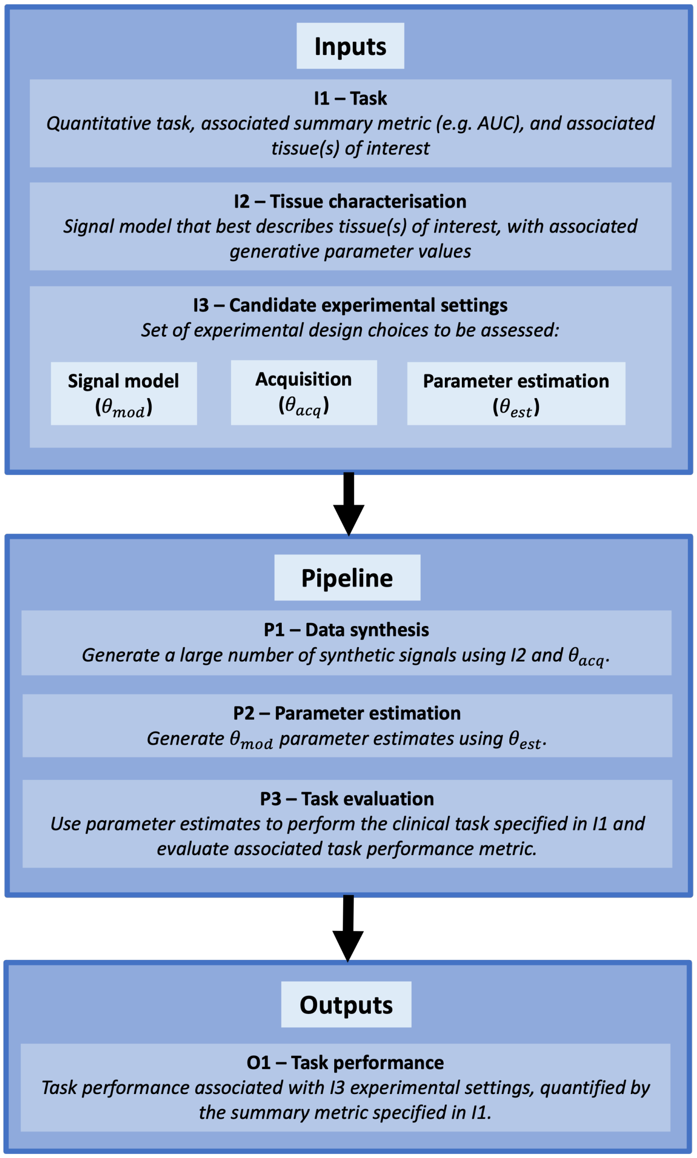
Inputs
I1 – Task
Quantitative task, associated summary metric (e.g. AUC), and associated tissue(s) of interest
I2 – Tissue characterisation
Signal model that best describes tissue(s) of interest, with associated generative parameter values
I3 – Candidate experimental settings
Set of experimental design choices to be assessed:
Signal model (θ_{mod})
Acquisition (θ_{acq})
Parameter estimation (θ_{est})
Pipeline
P1 – Data synthesis
Generate a large number of synthetic signals using I2 and θ_{acq}.
P2 – Parameter estimation
Generate θ_{mod} parameter estimates using θ_{est}.
P3 – Task evaluation
Use parameter estimates to perform the clinical task specified in I1 and evaluate associated task performance metric.
Outputs
O1 – Task performance
Task performance associated with I3 experimental settings, quantified by the summary metric specified in I1.

Figure 2-6: Graphical overview of proposed CED assessment pipeline.

I3: Candidate experimental design choices.
Example: qMRI acquisition protocol, qMRI model (may differ from the 'ground-truth' model in I2), parameter(s) selected for classification, parameter-estimation method, SNR (from echo time, repetition time, number of repetitions, average size of ROI), etc.

Simulation

P1: Data synthesis.
Example: A large number, sufficient to reduce sampling errors (e.g. 10,000),

of qMRI signals are synthesised according to I2 for each tissue type (healthy, diseased), using the designated model, drawing from the associated parameter distribution. Each qMRI signal is sampled and corrupted with Rician noise as detailed in I3.

P2: Parameter estimation.
Example: Each sampled noisy signal is analysed using the specified fitting method (I3) to generate parameter estimates.

P3: Task evaluation.
Example: The task (I1) is evaluated for the complete set of parameter estimates (P2), and an ROC curve is generated.

Output

O1: Task performance.
Example: AUC computed from ROC curves output as summary metric of I1 task performance associated with I3 experimental design and I2 tissue.

In this way, for a given I1-I3, the pipeline outputs a prediction of associated task performance. If a wide range of competing candidate experimental settings are assessed, the pipeline's task performance predictions can be used to select the optimal experimental design.

2.4.2 Implementation

The proposed pipeline was implemented in MATLAB 2019a (*The Math Works Inc.*) and applied to two exemplar qMRI experiments (E1 & E2). These experiments consisted of *subtyping*: classifying tissues as belonging to one of two disease stages. The task performance associated with such experiments is the ability to correctly classify unknown tissue.

E1 and E2 were chosen to correspond to different tasks within the same tissue, so as to demonstrate the advantages the pipeline offers over the task-agnostic status quo. For both experiments, a range of candidate θ_{mod} and θ_{est} were proposed and associated task performance assessed. Any situation where the optimal choice of model or fitting method is found to be task-specific represents a failure of current task-agnostic CED approaches.

Clinical context

Spondyloarthritis (SpA) was chosen as the clinical context for E1 and E2, as dMRI is increasingly being used as a tool for assessing its disease state [71, 72, 73]. SpA, an inflammatory disease which affects the bone and joints, is characterised by a range of abnormalities which include subchondral bone marrow oedema near the sacroiliac joint (SIJ) margins [74]. These inflammatory lesions, which can be subtyped as either 'active' or 'chronic', have been found to be well-described by the intravoxel incoherent motion (IVIM) dMRI model[5], with IVIM parameter estimates being sensitive to changes in pathology [50, 73].

Methods

Both E1 and E2 involve classifying tissues as belonging to one of two SpA subtypes; E1 differentiates healthy tissue from 'chronic' lesions, whereas E2 compares 'chronic' lesions to 'active' ones.

Synthetic qMRI data corresponding to each subtype was generated (θ_{acq}) and signal model (θ_{mod}) parameters were estimated (θ_{est}). The resulting parameter estimates were used to classify the tissue. This classification was assessed with ROC curves and associated AUC. Data was synthesised from the IVIM model [75]:

$$\frac{S(b)}{S_0} = fe^{-b(D_{fast}+D_{slow})} + (1-f)e^{-bD_{slow}} \tag{2.10}$$

where $S(b)$ is the MRI signal at diffusion weighting b, S_0 is the signal at $b=0$, f is the perfusion fraction, D_{fast} is the pseudo-diffusivity of perfusing water, and D_{slow} is the diffusivity of non-perfusing water.

Experimental settings are shown in Table 2.1: to simplify analysis of results, tissue parameters were chosen such that only *one* IVIM parameter varied between subtypes per task.

Dataset	**Task**	**Tissue**	**Generative model**	f	D_{slow} $(10^{-3}\, mm^2/s)$	D_{fast} $(10^{-3}\, mm^2/s)$	**Effective SNR**	**Sampling** (s/mm^2)	**Model & fitting method**
Simulated	E1	Healthy	Intravoxel incoherent motion (IVIM)	0.09	0.35	123	20	0, 10, 20, 40, 80, 100, 200, 400, 600	IVIM: sNLLS & bcNLLS ADC: wLS
		Chronic		0.12	0.35	123			
	E2	Active		0.12	0.60	123			
		Chronic		0.12	0.46	123			

Table 2.1: Experimental settings. Synthetic signals were generated from the IVIM model. Both IVIM and ADC parameters were estimated from IVIM-synthesised data.

[5] See Appendix C.

Rician noise was added at $SNR = 20$, defined as:

$$S_{noisy}(b) = \sqrt{\mathcal{N}(S_{noisefree}(b), \sigma^2)^2 + \mathcal{N}(0, \sigma^2)^2}; \quad SNR = \frac{S_0}{\sigma} \tag{2.11}$$

where $S_{noisefree}$ is the noise-free IVIM signal, S_0 is the noise-free signal at $b = 0$, and $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution of mean μ and standard deviation σ.

Data was sampled at 9 b-values representative of typical SpA dMRI acquisition. Two models were fit to this data: IVIM and ADC.

IVIM fitting

The signal was normalised by S_0 and IVIM fitting was performed with either segmented non-linear least squares (sNLLS) [70] or bound-constrained non-linear least squares (bcNLLS) [70]: $f \in [0, 1]$, $D_{fast} \in [0, 500]10^{-3}mm^2/s$, $D_{slow} \in [0, 10]10^{-3}mm^2/s$; fitting was seeded with mean parameter values across tissues within each dataset to mimic real-world experiments in which individual patient classification is not known. In the case of two-step sNLLS fitting, $b_{threshold}$ is chosen in light of expected tissue properties; in this instance, $b_{threshold} = 50s/mm^2$.

ADC fitting

Fitting was was also performed using the ADC model, defined as:

$$\frac{S(b)}{S_0} = e^{-bADC} \tag{2.12}$$

where $S(b)$ is the MRI signal at diffusion weighting b, S_0 is the signal at $b = 0$, and ADC is the apparent diffusion coefficient.

This model was fit on the log-transformed signal using weighted least-squares (WLS) linear regression [8], a single-shot non-iterative method for obtaining maximum likelihood estimates.

2.4.3 Results

Figure 2-7 shows the advantages the pipeline offers over existing CED assessment methods. It demonstrates that, within a single disease, different classification tasks may be best served by different (a) models and (b) model fitting methods.

Regarding model selection, both ADC and IVIM models are sensitive to changes in tissue microstructure ($AUC > 0.5$) in both E1 and E2. Within E1, where tissues

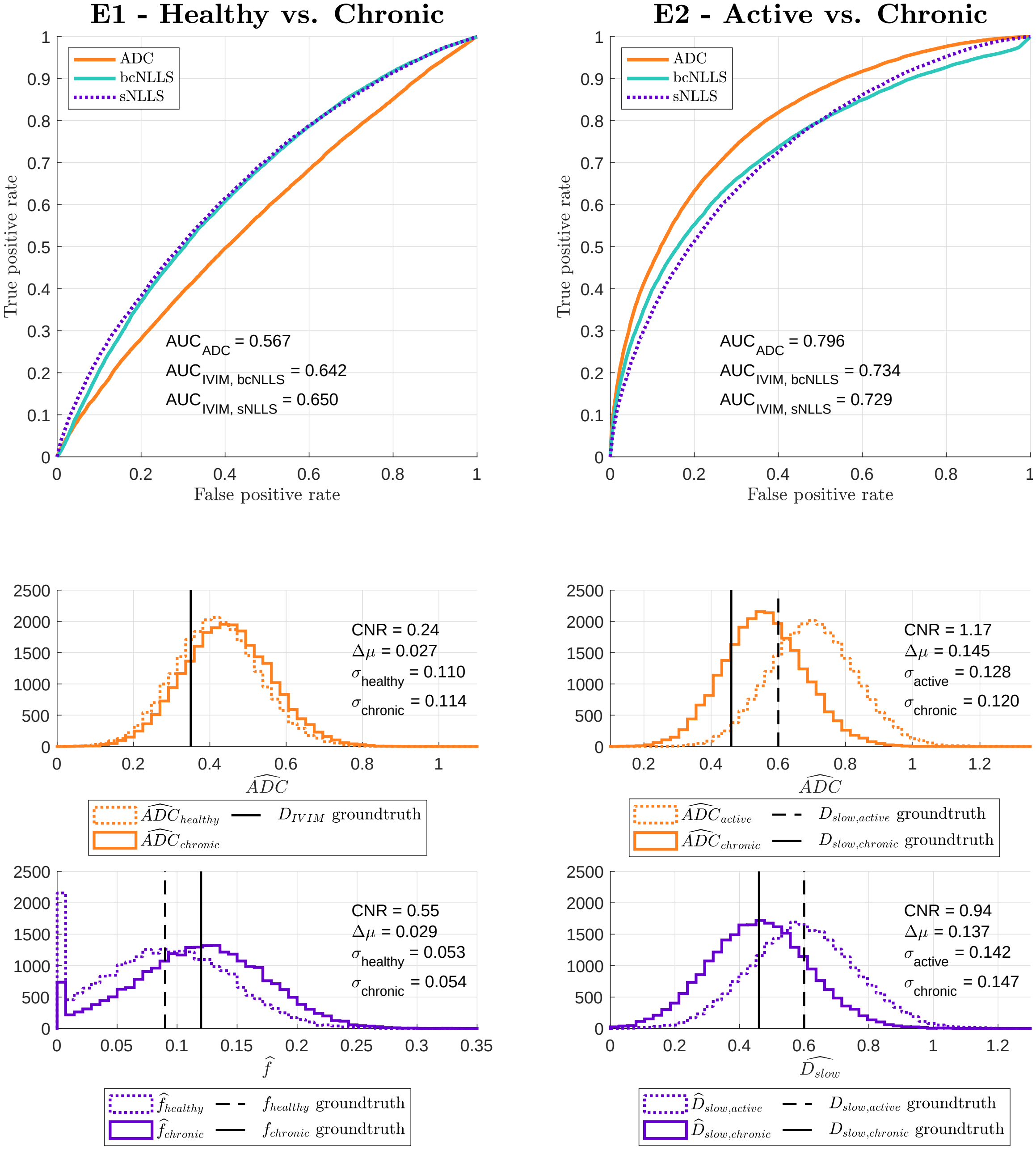

Figure 2-7: ROC curves and associated AUC values for two exemplar classification tasks, together with the distributions of parameter estimates underlying these ROC curves. Despite IVIM being the generative model for both tasks, it is outperformed by ADC in Task E2. Within IVIM, sNLLS outperforms bcNLLS in Task E1; the opposite is true in Task E2.

differ by their perfusion fraction, ADC is less sensitive to population differences than IVIM ($AUC_{ADC} < AUC_{IVIM}, CNR_{AUC} < CNR_{IVIM}$). In contrast, in E2, where signal differences arise from variation in D_{slow}, the biased ADC model *outperforms* the ground truth generative model (IVIM). ADC, being the ‘incorrect’ model, gives upward-biased estimates of diffusivity, but results in improved classification performance ($AUC_{ADC} > AUC_{IVIM}$), due primarily to its lower parameter estimation variance ($\sigma^2_{ADC} < \sigma^2_{IVIM}$). The upward bias does not adversely affect task performance as it is consistent across tissue types. These results demonstrate that (a) in general, biased models may outperform unbiased ones and (b) optimal model selection may vary, within a single tissue type, depending on the task of interest; model selection should not be task-agnostic.

With regard to parameter estimation, the optimal IVIM fitting method varies slightly between tasks (sNLLS for E1 and bcNLLS for E2). This result arises from the fact that fitting method performance varies across model parameters; the best method for estimating f (as needed in E1) may differ from that for estimating D_{slow} (as per E2). Figure 2-8 reveals the source of these differences.

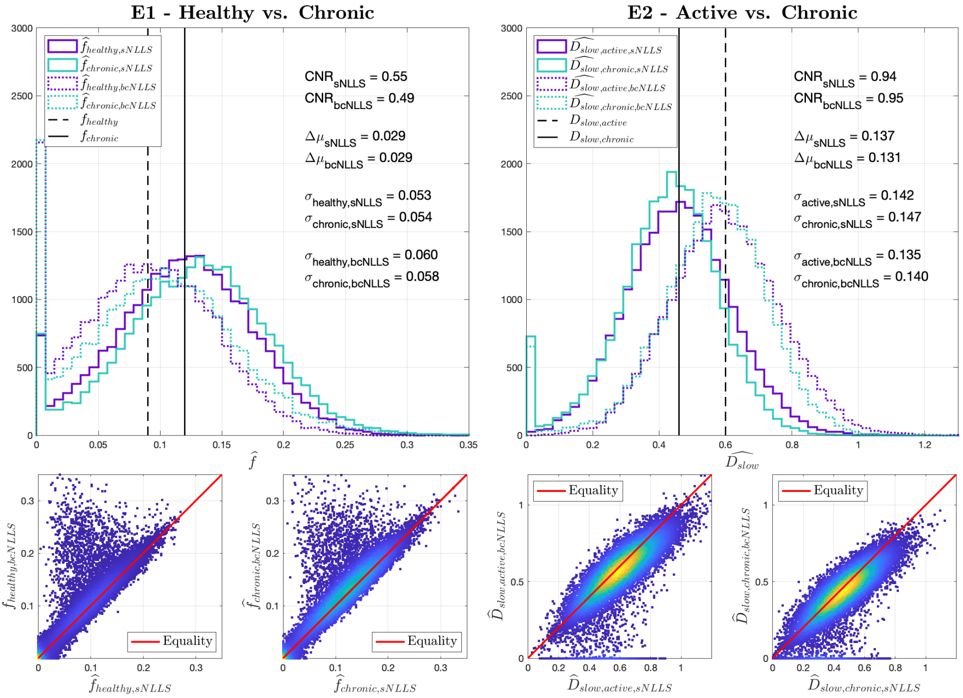

Figure 2-8: Comparison of sNLLS and bcNLLS fitting methods within tasks E1 and E2. The top panels show the distributions of IVIM parameter estimates; the bottom row shows the correlation between sNLLS and bcNLLS parameter estimates. The difference in fitting method performance is most pronounced in E1, where bcNLLS results in on-average higher f than sNLLS.

2.5 Discussion and conclusions

This chapter has introduced the notion of P_{target} and ϕ, the information-rich parameter estimates produced by an 'ideal' qMRI experiment. This has enabled an analysis of the CED status quo: setting $P_{target} = T$ and, in so doing, the prioritization of *absolute information* at the expense of *relative contrast*.

The central argument of this chapter is that this choice of P_{target} is inappropriate for qMRI contexts, such as clinical classification, where outcome is measured in terms of effect size rather than measurement accuracy. In such settings, P_{target} should be adjusted to maximise task performance, rather than minimise measurement error. This mimics real-world task-driven experimental design, where clinical data is acquired and used to assess experimental settings.

To this end, this chapter has proposed and implemented a computational pipeline for assessing experimental designs on their task performance, rather than parameter estimation, capabilities. This method addresses limitations of current approaches to CED assessment, which rely on task-agnostic measures of parameter-estimation accuracy or precision, thereby providing indirect, potentially unreliable, predictions of task sensitivity and specificity. In contrast, by explicitly simulating the interactions between myriad experimental design choices, the proposed method is able to directly predict task performance.

Two exemplar dMRI experiments (E1 and E2) were used to demonstrate that traditional parameter-estimation CED metrics lead to sub-optimal dMRI task performance. The fact that optimal model selection and fitting method varies across tasks, within pathology, is evidence of a failure of current practice. Assessed on parameter-estimation performance alone, IVIM outperforms ADC; yet, in Task E2, ADC yields higher task performance. Assessed on parameter-estimation performance, sNLLS outperforms bcNLLS [67, 66]; yet, in Task E2, bcNLLS is the optimal IVIM parameter estimation algorithm. Task-agnostic parameter-estimation metrics, such as bias or variance, are thus shown to be unreliable predictors of task performance. Task-specific assessment, which directly measures experimental outcome, is required to reliably assess experimental designs. Such assessment is inaccessible within current CED practice.

2.5.1 Relation to existing work

There is at present no similar CED framework for task-driven assessment. The closest to this approach is a method of MRI protocol optimisation driven by statistical decision theory [76]. The authors argue for a task-driven approach to protocol optimisation. However, their approach differs from the one presented here in an important way. Theirs maximises task performance via acquired MRI signals, which provide an indirect measure of the underlying difference in tissue properties. In contrast, the newly-proposed approach assesses task performance by directly considering the properties of interest: qMRI parameter estimates. Since tissue properties are estimated from measured MRI signals by means of model fitting, the choice of model and fitting methodology impacts task performance and is therefore explicitly assessed by the proposed method.

The proposed pipeline mimics, in-silico, the gold-standard method used to assess task-driven experimental designs: acquiring rich, super-sampled clinical datasets which are successively sub-sampled, with each reduced dataset assessed on its associated task performance [67, 70]. These methods are data-intensive and may be impractical for many applications. The proposed framework offers a computational alternative that (i) makes task-driven experimental assessment more accessible and (ii) can be used to narrow the search space of experimental design choices before real data is required, thereby informing, focusing, and substantially shortening, any subsequent task-driven clinical evaluation and validation.

2.5.2 Use-cases and broader scope

The proposed method produces assessment metrics for experimental designs, and it is left to the end-user to decide what to do with these metrics. The simplest use-case is experimental design selection: a range of plausible experimental settings are assessed, and the task-optimal experimental design is chosen from this set. Another use-case is optimisation, whether manual or automated: experimental settings are repeatedly adjusted and assessed; changes that improve task performance are retained, leading to iterative optimisation.

Regardless of use-case, the framework can be applied to a broad range of quantitative MRI contexts; the pipeline is compatible with any quantitative model-based task-driven application (e.g. fat fraction mapping [77]) for which a quantitative task (I1), well-characterised tissue properties (I2), and the ability to generate synthetic signal (I3) are available.

Furthermore, E1 and E2 have used AUC as a summary task-performance metric due to the availability of clinical values to validate against. However, the framework's intermediate output - a distribution of parameter estimates, used here to generate ROC curves for binary classification - mimics 'real' qMRI outputs and can be assessed as such, regardless of qMRI task.

2.5.3 Limitations and next steps

One limitation of the proposed approach is that it requires clear knowledge of the relationship between tissue properties and the underlying pathology of interest (i.e. I2). Whilst this knowledge may not currently exist for all clinical use-cases, it is a natural by-product of ongoing basic research; the pipeline provides a proof-of-concept computational framework to exploit these relationships.

Another potential limitation is the reliance on simulation. Simulation approximates the biophysical processes that underpin qMRI data acquisition. It is unclear how reliable the pipeline's task-performance predictions are: can they be trusted to match those that would be obtained by combining θ_{acq}, θ_{mod}, and θ_{est} and performing the qMRI experiment in-vivo? This question is answered in Chapter 3.

Chapter 3

Validating the method: in-silico vs. in-vivo

Contents

This chapter is adapted from:

- **Epstein SC**, Bray TJP, Hall-Craggs MA, Zhang H, *Task-driven assessment of experimental designs in diffusion MRI: a computational framework*, 2021, PLOS ONE 16(10): e0258442
- **Epstein SC**, Bray TJP, Hall-Craggs M and Zhang H, *Towards a computational framework for task-driven experimental design*, 2021, ISMRM Annual Meeting 2021

3.1 Validation of proof-of-concept

In the previous chapter, a case was made for reframing qMRI CED in terms of task performance maximisation, and a proof-of-concept computational implementation (the ‘pipeline’) was presented. This chapter takes this implementation one step further, and assesses its value in real clinical contexts.

As discussed in Chapter 1, non-computational experimental design assesses experimental designs by implementing them: acquiring data, testing it, and making an empirical judgement on its quality. CED approaches mimic some (or all) aspects of this process in-silico and, in so doing, promise a reduction in costly data acquisition; the pipeline proposed in Chapter 2 does this mimicry holistically and in a task-driven manner.

For this mimicry to be useful, it must replace the 'real thing': going out and acquiring real data. In other words, the in-silico outputs (task performance predictions) must match those that would be obtained from in-vivo experiments. This chapter describes validation experiments which test this correspondence by directly comparing the proposed pipeline's output to associated real-world task performance metrics.

3.1.1 Implementation

Two qMRI datasets were identified for which performance on a classification task was either published or could be computed. For each dataset, the proposed pipeline mirrored real-world experimental design choices (θ_{acq}, θ_{mod}, θ_{est}) and was used to make in-silico predictions of task performance. These predictions, made without access to any acquired data, were compared against the in-vivo task performance metrics. Agreement between prediction and in-vivo observation validates the pipeline's usefulness in qMRI CED assessment.

Clinical context

For consistency with Chapter 2, SpA-related validation datasets were selected.

Methods

The first dataset ('Zhao') consists of 41 patients split across three SpA subtypes [73] and reported ROC/AUC curves for three classifcation tasks (V1-V3). The second dataset ('Bray') is comprised of 28 patients split across two SpA subtypes [50] and analysed for one task (V4). The four classification tasks represented by the two datasets are described in Table 3.1.

Dataset	**Task**	**Tissue**	**Generative model**	f	D_{slow} $(10^{-3}\,mm^2/s)$	D_{fast} $(10^{-3}\,mm^2/s)$	**Effective SNR**	**Sampling** (s/mm^2)	**Model & fitting method**
Zhao	V1	Chronic	Intravoxel incoherent motion (IVIM)	0.12 ± 0.02	0.35 ± 0.11	124.7 ± 13.7	150.6	0, 10, 20, 30, 50, 80, 100, 200, 400, 800	IVIM: sNLLS
		Healthy		0.09 ± 0.02	0.34 ± 0.09	122.7 ± 18.3			
	V2	Active		0.12 ± 0.03	0.99 ± 0.39	123.9 ± 19.9			
		Chronic		0.12 ± 0.02	0.35 ± 0.11	124.7 ± 13.68			
	V3	Active		0.12 ± 0.03	0.99 ± 0.39	123.9 ± 19.9			
		Healthy		0.09 ± 0.02	0.34 ± 0.09	122.7 ± 18.3			
Bray	V4	Inflamed		0.07 ± 0.08	1.91 ± 0.56	24.2 ± 28.5	56.3	0, 50, 100, 300, 600	IVIM: bcNLLS ADC: wLS
		Normal		0.05 ± 0.04	0.92 ± 0.26	44.6 ± 35.2			

Table 3.1: Computational pipeline settings for V1-4. Synthetic signals were generated using parameters drawn from normal distributions taken from Zhao [73] or Bray [50]. Effective SNR for V4 was calculated from Bray's $b = 0$ images, and, for Tasks V1-3, adjusted by mean ROI size and acquisition differences (echo time, voxel size, number of repetitions, etc.) between Bray and Zhao's experiments, reported [50] and [73] respectively. In V4, ADC values were estimated from IVIM-synthesised data.

Synthetic qMRI data corresponding to each subtype were generated, from the IVIM model, using in-vivo-matched acquisition protocols (θ_{acq}) and signal model (θ_{mod}) parameters were estimated (θ_{est}). For each validation task, the resulting parameter estimates were used to classify the tissue. This classification was assessed with ROC curves and associated AUCs.

IVIM tissue parameters were drawn from Gaussian distributions representing SpA lesions relevant to each classification task, and Rician noise (as defined in Equation 2.11) was added at SNRs commensurate with each real-world acquisition.

As before, both the IVIM and ADC models were fit to the data using the techniques described in Section 2.4.2. In the case of two-step sNLLS fitting, $b_{threshold}$ should be adjusted based on expected tissue properties; in this instance, we replicated Zhao's choice of $b_{threshold} = 200s/mm^2$.

The in-silico ROC/AUC predictions were validated across multiple microstructural models (IVIM, ADC) and model parameters (D_{fast}, D_{slow}, f, ADC) in two ways. Firstly, large patient populations (1000 x clinical dataset size) were simulated to obtain numerically robust task performance predictions. Secondly, to quantify agreement between these results and the smaller clinical datasets, these large datasets were repeatedly sub-sampled, each time matching real-world patient numbers. The resulting distribution of sub-sampled task performance metrics was compared to the in-vivo AUC.

3.1.2 Results

Figure 3-1 compares V1-V3 task performance predictions to ground-truth clinical ROC curves, and shows that our pipeline accurately predicts (a) the qualitative form of the ROC curves, (b) the relative task performance of different experimental settings, and (c) the absolute task performance (AUC) of each experimental design.

These findings are replicated for Task V4 in Figure 3-2.

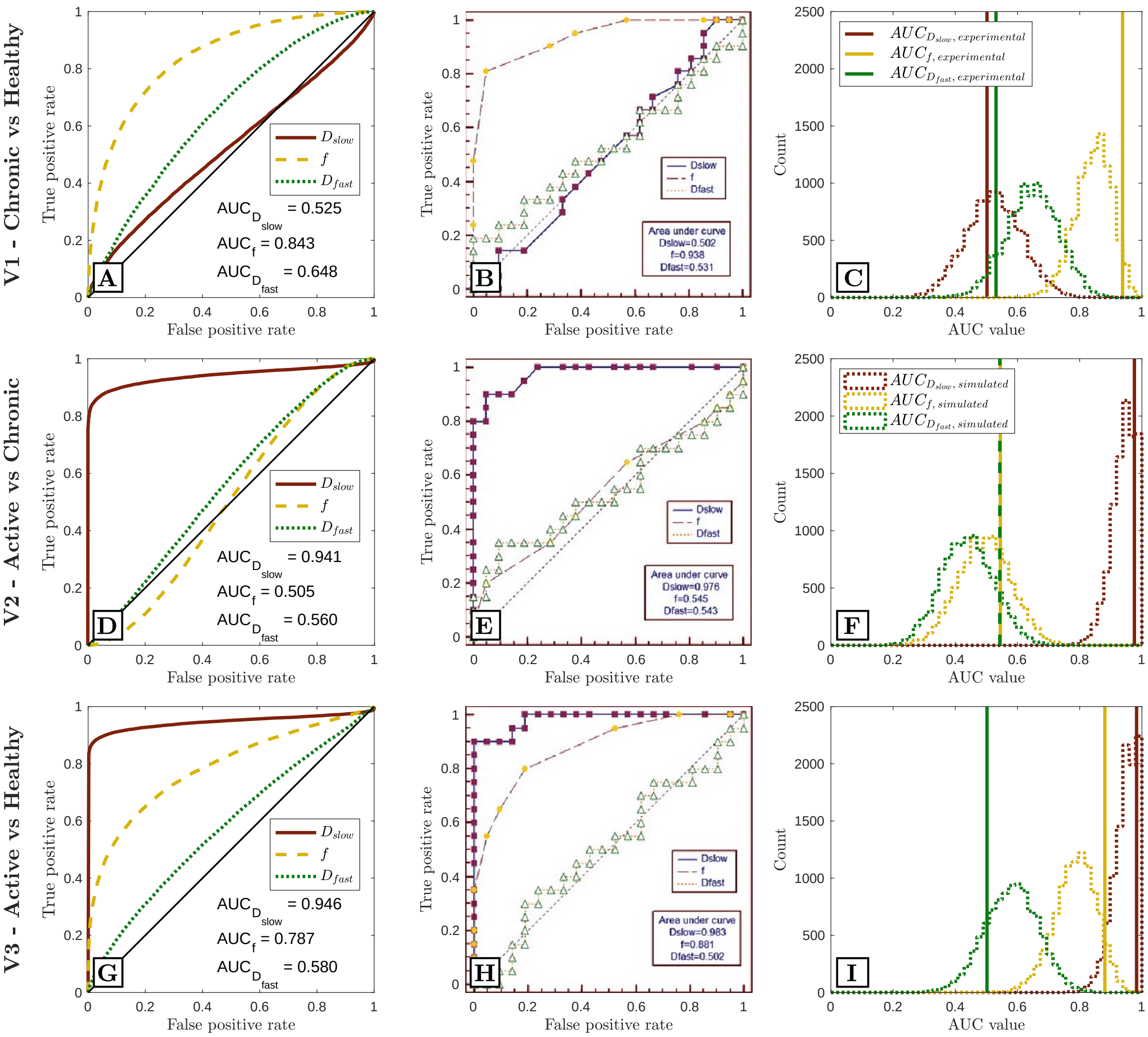

Figure 3-1: Simulated (first column) vs. clinical (middle column) ROC curves for Zhao's dataset (V1-V3). The third column compares clinical AUC values (vertical lines) to simulated AUC values (distributions) when sub-sampling simulated data to match clinical study sizes. All ROC curves are qualitatively similar; the relative performances (AUC values) of different IVIM parameters are equal; all AUC values are in numerical agreement once clinical sample size is considered.

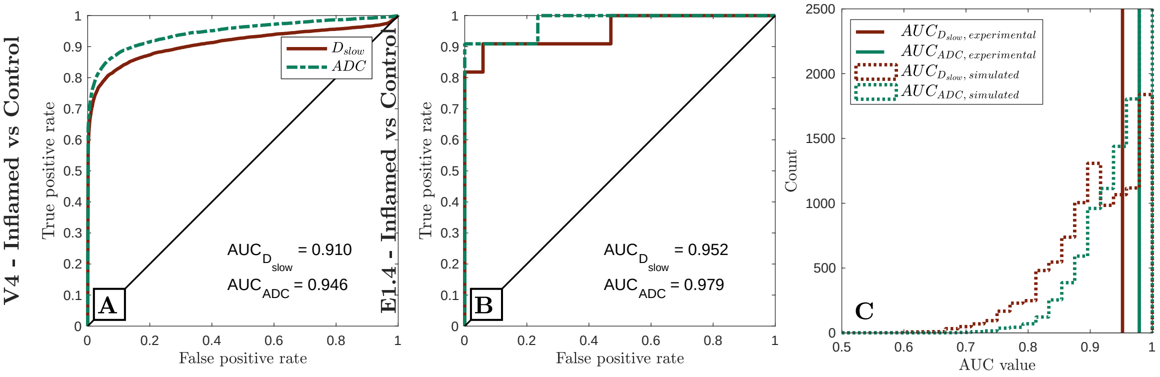

Figure 3-2: Simulated (first column) vs clinical (middle column) ROC curves for Bray's dataset (V4). The third column compares clinical AUC values (vertical lines) to simulated AUC values (distributions) when sub-sampling simulated data to match clinical study sizes. As in Figure 3-1, ROC curves are qualitatively similar; the relative performance (AUC values) of dMRI models are equal; all AUC values are in numerical agreement once clinical sample size is accounted for.

3.2 Discussion and conclusions

This chapter validates the task-performance predictions of the computational pipeline proposed in Chapter 2 by comparing them to those obtained in-vivo.

Four validation experiments (V1-V4) were extracted from two datasets, and the task performance associated with experimental choices (θ_{mod}, θ_{est}, and θ_{acq}) were predicted and compared to those obtained in-vivo. Qualitative and quantitative agreement across all validation experiments shows that the proposed pipeline's task-performance predictions are accurate and reliable. This has three consequences. Firstly, the pipeline can be used to differentiate, between a set of candidate experimental settings, the single experimental design which maximises task performance. Secondly, it can be used to iteratively adjust experimental design parameters, such as acquisition time, until specific, clinically-required task performance is achieved. Thirdly, it validates the results of E1 and E2 in Chapter 2, which demonstrated the benefits of task-driven CED approaches.

3.2.1 Implications & outlook

As described in Chapter 2, the pipeline simply assesses candidate experimental designs. In its pre-validation proof-of-concept state, it was presented as a tool for experimental design selection or iterative optimisation. The validation results presented here enable a third use-case: calculating acquisition-time requirements. The fact that task-performance predictions are quantitatively accurate mean that the θ_{acq} required for a specified task performance (e.g. $<$10% false-positive-rate) could be determined in-silico. Such calculations are not possible with existing task-agnostic assessment methods.

The proposed in-silico task-specific assessment promises improved experimental design selection and optimisation: all experimental choices, from data acquisition to analysis, can be analysed and compared without the need for expensive, time-consuming data acquisition. Model fitting methods can be assessed in a more meaningful, task-specific manner. Traditionally unfavoured, high-bias models, can be considered, assessed, and selected.

Although the pipeline can simply replace the assessment stage in current CED practice, it naturally lends itself to being incorporated into an overarching task-driven CED framework: combining the proposal of candidate experimental designs with computational optimisation, using the presented work as an accurate, task-specific optimisation metric.

3.2.2 Next steps

So far, this thesis has (i) proposed a re-framing of the purpose of qMRI experiments, away from tissue measurement and towards task performance, (ii) developed a corresponding proof-of-concept CED pipeline, (iii) validated the associated task performance predictions, and, in so doing, (iv) presented a demonstrably-useful in-silico framework for making task-driven qMRI experimental design choices.

Now that we have a tool to make optimal choices, this thesis turns to *improving the choices themselves*. Rather than focus on θ_{acq} (protocol optimisation) or θ_{mod} (model development), Chapter 4 analyses θ_{est}, the methods by which $\overline{P}$ are extracted from $S_{acquired}$. What follows is an analysis of existing parameter estimation methods, and a proposal for a novel approach which slots naturally into the task-driven pipeline described thus far.

Chapter 4

Improving parameter estimation: a novel deep learning method

Contents

This chapter is adapted from:

- **Epstein SC**, Bray TJP, Hall-Craggs MA and Zhang H, *Choice of training label matters: how to best use deep learning for quantitative MRI parameter estimation*, 2022, arXiv:2205.05587

- **Epstein SC**, Bray TJP, Hall-Craggs MA and Zhang H, *Quantitative MRI parameter estimation with supervised deep learning: MLE-derived labels outperform groundtruth labels*, 2022, ISMRM Annual Meeting 2022

qMRI promises many advantages over its conventional imaging counterpart: increased sensitivity, specificity, reproducibility, interpretabilty, and tissue insight. And yet, conventional MRI remains more popular in clinical contexts.

A significant barrier to widepsread qMRI adoption is the increased acquisition and post-processing cost. One of the largest time and resource bottlenecks in post-processing is parameter estimation: fitting the qMRI signal model M to the acquired data. Traditionally, each voxel has required its own independent model fit: solving for the parameters that best described the single voxel's data. The computational cost of this curve-fitting process, which scales with both voxel number and model complexity, has become a bottleneck for modern qMRI experiments.

Accelerating curve fitting with deep learning (DL) was first proposed more than 30 years ago [78], but has only recently gained popularity within the qMRI community [60, 61, 58, 59, 79]. Just like traditional methods, DL relies on model fitting, but the model being fitted is a fundamentally different one. Instead of fitting a qMRI *signal model* to a single voxel of interest (i.e. curve fitting), DL methods fit ('train') a deep neural network (DNN) model to an ensemble of training voxels. This model maps a single voxel's signal to its corresponding qMRI parameters; the unknowns in its fitting are network weights, rather than qMRI parameters. Once this DNN model has been fitted to ('trained on') the training data, parameter estimation is reduced to simply applying it to new unseen data, one voxel at a time. This approach offers two broad advantages over traditional fitting: (i) computational cost is amortised: despite being more computationally expensive than one-voxel signal model fitting, DL training only needs to be performed once, for any number of voxels; once trained, networks provide near-instantaneous parameter estimates on new data, and (ii) computational cost is front-loaded: model training can be performed away from the clinic, before patient data is acquired.

To date, most DL qMRI fitting methods have been implemented within a supervised learning framework [60, 80, 58, 59, 79, 81, 82, 83]. This approach trains DNNs to predict groundtruth qMRI model parameters from noisy qMRI signals. When compared to conventional fitting, this approach has been found to produce high bias, low variance parameter estimates [84, 83].

An alternative class of DL methods has also been proposed, sometimes referred to as unsupervised learning [61, 85], but more accurately described as self-supervised [86]. In this framework, training labels are not explicitly provided, but are instead extracted by the network from its training input. This label generation is designed such that the network learns to predict signal model parameters corresponding to noise-free signals that most-closely approximate noisy inputs. This self-supervised approach has been found to produce similar results to conventional non-DL fitting, i.e. lower bias and higher variance than its groundtruth-labelled supervised alternative [61, 84].

From an information theoretic standpoint, the comparison between supervised and self-supervised performance raises an obvious still unanswered question. How can it be that supervised methods, which provide strictly more information during training than their self-supervised counterparts, produce more biased parameter estimates?

This chapter answers this question by showing that this apparent limitation of supervised approaches stems purely from the selection of groundtruth training labels. By training on labels which are deliberately not groundtruth, this work shows that the low-bias parameter estimation previously associated with self-supervised methods can be replicated – and improved on – within a supervised learning framework.

This approach sets the stage for a single, unifying, deep learning parameter estimation framework, based on supervised learning, where trade-offs between bias and variance can be made, on a task-specific basis, by careful adjustment of training label.

4.1 Existing parameter estimation methods

Before introducing the novel parameter-estimation method which underpins this framework, it is necessary to describe the existing methods it complements. In keeping with machine learning conventions, the mathematical notation used in this chapter differs from that used in Chapters 2 and 3. Within this notation, qMRI produces quantitative spatial maps by extracting biomarkers $\overline{y}$ from MR data x.

4.1.1 Conventional fitting (MLE)

This method, described in detail in Section 2.1, extracts biomarkers by performing a voxelwise model fit every time new data is acquired. An appropriate signal model M is required, parameterised by n_y parameters of interest; for each combination of y, the probability of observing the acquired data x subject to noise model ϵ is known as the likelihood $\mathcal{L}$. The model parameters $\overline{y}$ which *maximise* $\mathcal{L}$ are assumed to best represent the tissue contained within the voxel of interest:

$$\overline{y} = \arg\max_{y} \mathcal{L}(x, z|y, \epsilon) \tag{4.1}$$

for sampling scheme z[1]. This optimisation can be equivalently expressed as a *minimisation* of some loss function L_{MLE}:

$$\overline{y} = \arg\min_{y} L_{MLE}(x, z|y, \epsilon) \tag{4.2}$$

Each of these minimisations has n_y unknowns, which are solved independently across different voxels; the computational cost scales linearly with the number of

[1] Note that within the CED notation used throughout this Thesis, z is subsumed by θ_{acq}.

voxels n_v.

Developments in qMRI acquisition and analysis have led to increased (a) image spatial resolution (i.e. greater n_v) and (b) model complexity (i.e. greater n_y), such that conventional MLE fitting has become increasingly computationally expensive.

4.1.2 Existing deep learning methods

DL approaches address this by reframing n_v independent problems into a *single* global model fit: learning the function $\mathcal{F}$ that maps any x to its corresponding groundtruth y_{gt}:

$$y_{gt} = \mathcal{F}(x) \tag{4.3}$$

Deep neural networks aim to approximate this function by composing a large but finite number of building-block functions[2], parametrised by n_w network parameters w ('weights'):

$$y = \mathcal{F}_{approx}(x|w) \tag{4.4}$$

In this context, model fitting ('training'), is performed over network parameters w and involves minimising $\mathcal{F}_{approx}$'s mean training loss L over a large set of training examples $\mathbb{X}$; the trained network is defined by the best-fit parameters $\overline{w}$.

$$\overline{w} = \arg\min_{w} L(\mathcal{F}_{approx}(x|w), \mathbb{X}) \tag{4.5}$$

where the best-fit function $\overline{\mathcal{F}}$ becomes:

$$\overline{\mathcal{F}}_{approx} = \mathcal{F}_{approx}(x|\overline{w}) \tag{4.6}$$

Such that qMRI parameter estimates are obtained from:

$$\overline{y} = \overline{\mathcal{F}}_{approx}(x) \tag{4.7}$$

The fitting problem described by Equation 4.5, whilst more computationally expensive to solve than any individual voxel ($n_v = 1$) MLE, is only tackled once; once

[2] See Appendix D for a brief introduction to DNNs.

$\overline{\mathcal{F}}_{approx}$ is learnt, it can be applied at negligible cost to new, unseen, data. This promise of rapid, zero-cost parameter estimation has led to the development of two broad classes of DL-based parameter estimation methods.

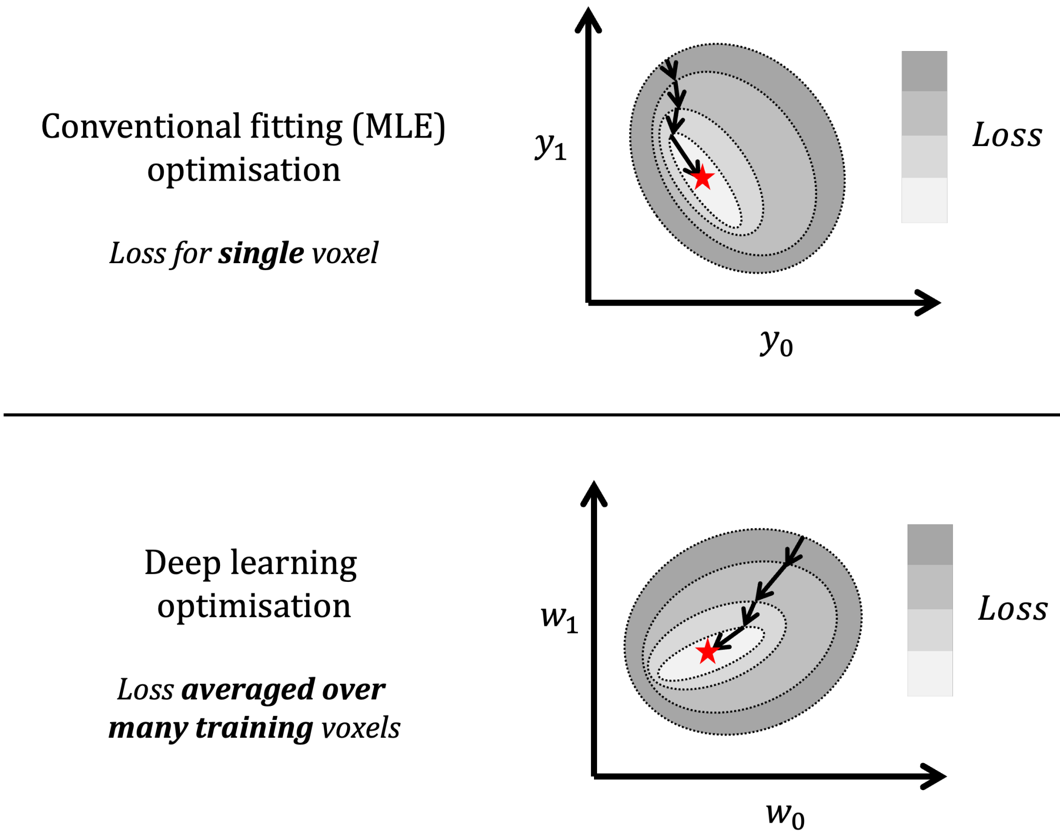
Conventional fitting (MLE) optimisation
Loss for single voxel
y_1
y_0
Loss
Deep learning optimisation
Loss averaged over many training voxels
w_1
w_0
Loss

Figure 4-1: Graphical comparison between conventional fitting (MLE) and deep learning parameter estimation, showing a 2D simplification of high-dimensional optimisation process. The black arrows represent iterative steps towards the minimum of a loss functions, shown as a red star. MLE optimises a loss (defined in qMRI parameter space Y) independently for each acquired voxels. Deep learning optimises a loss (defined in 'weight'-space W) averaged over all training voxels.

Supervised$_{GT}$ methods approximate $\mathcal{F}$ by defining the training loss L_{GT} as the difference between noise-free labels (groundtruth parameter values) and network outputs (noise-free parameter estimates), calculated in the parameter space Y:

$$\text{Supervised}_{\text{GT}} \text{ training loss} \equiv L_{GT} = \sum_{i=1}^{n_{train}} \|\mathcal{W} \cdot (\overline{y}_i - y_{gt,i})\|^2 \tag{4.8}$$

where n_{train} is the number of training samples and $\mathcal{W}$ is a tunable weight matrix which accounts for magnitude differences in signal model parameters. $\mathcal{W}$ is generally a diagonal matrix, with each diagonal element $\mathcal{W}_{ii}$ corresponding to the relative weighting of qMRI parameter y_i; setting $\mathcal{W}$ as the identity matrix equally weights all parameters in the training loss.

These methods produce higher bias, lower variance parameter estimation than conventional MLE fitting [83, 84] and, by adjusting $\mathcal{W}$, can be tailored to selectively boost estimation performance on a subset of the parameter space Y.

In contrast, *Self-supervised* methods compute training loss L_{SS} within the signal space X, by minimising the difference between network inputs (noisy signals) and a filtered representation of network outputs (noise-free signal estimates):

$$\text{Self-supervised training loss} \equiv L_{SS} = \sum_{i=1}^{n_{train}} \|M(z|\overline{y}_i) - x_i\|^2 \tag{4.9}$$

where M is the qMRI signal model defined in Equation 2.2.

These methods, which perform similarly to conventional MLE fitting, produce lower bias, higher variance parameter estimation than *Supervised*$_{GT}$ [61, 84]. Unlike *Supervised*$_{GT}$, the relative loss weighting of different signal model parameters is limited by sampling scheme z.

Under Gaussian noise conditions, single-voxel *Self-supervised* loss (i.e. minimising the sum of squared differences between a noisy signal and its noise-free signal estimate) is indistinguishable from the corresponding objective function in conventional fitting.

In contrast, under the Rician noise conditions encountered in MRI acquisition, *Self-supervised* training loss no longer matches conventional fitting. Indeed, the sum of squared errors between noisy signals and noise-free estimates is not an accurate difference metric in the presence of Rician noise.

To summarise: existing supervised DL techniques are associated with high estimation bias, low variance, and end-user flexibility; in contrast, self-supervised methods have lower bias, higher variance, but are limited by the fact that their loss is calculated in the signal space X.

4.2 Proposed parameter estimation method

In light of the limitation of existing DL-based methods, this thesis proposes *Supervised*$_{MLE}$, a novel parameter estimation method which combines the advantages of *Supervised*$_{GT}$

and *Self-supervised* methods. This method is compared to existing techniques in Fig 4-2.

This method mimics *Self-supervised*'s low-bias performance by learning a regularised form of conventional MLE, but does so in the parameter space Y, within a supervised learning framework. This addresses the limitations of *Self-supervised*: Rician noise modelling is incorporated, and parameter loss weighting is not limited by sampling scheme z.

This method learns $\overline{\mathcal{F}}_{approx}$ by defining the training loss L_{MLE} as the difference between network predictions and *pre-computed conventional MLE labels.* These labels act as proxies for the groundtruth parameters we wish to estimate:

$$\text{Supervised}_{\text{MLE}} \text{ training loss} \equiv L_{MLE} = \sum_{i=1}^{n_{train}} \|W \cdot (\overline{y}_i - y_{MLE,i})\|^2 \tag{4.10}$$

The proposed method offers one final advantage over *Self-supervised* approaches. In addition to the parameter estimation improvements relating to noise model correction and parameter loss weighting, it naturally interfaces with *Supervised$_{GT}$*. In so doing, it presents the opportunity to combine low-bias and low-variance methods into a single, tunable hybrid approach, by a simple weighted sum of each method's loss function:

$$\text{Hybrid training loss } = \alpha L_{MLE} + (1-\alpha)L_{GT} \tag{4.11}$$

for $0 \leq \alpha \leq 1$.

4.3 Comparison to existing methods

The proposed method was trained and tested on both simulated and in-vivo data. Its parameter estimation performance was compared to the other methods described above, and a proof-of-concept 'hybrid' training loss was implemented and evaluated.

4.3.1 Implementation

Three classes of network were investigated and compared: *Supervised$_{GT}$*, *Self-supervised*, and *Supervised$_{MLE}$*, as described in Fig 4-2. Additionally, to control for differences in loss function weighting between supervised and unsupervised methods, *Self-supervised*

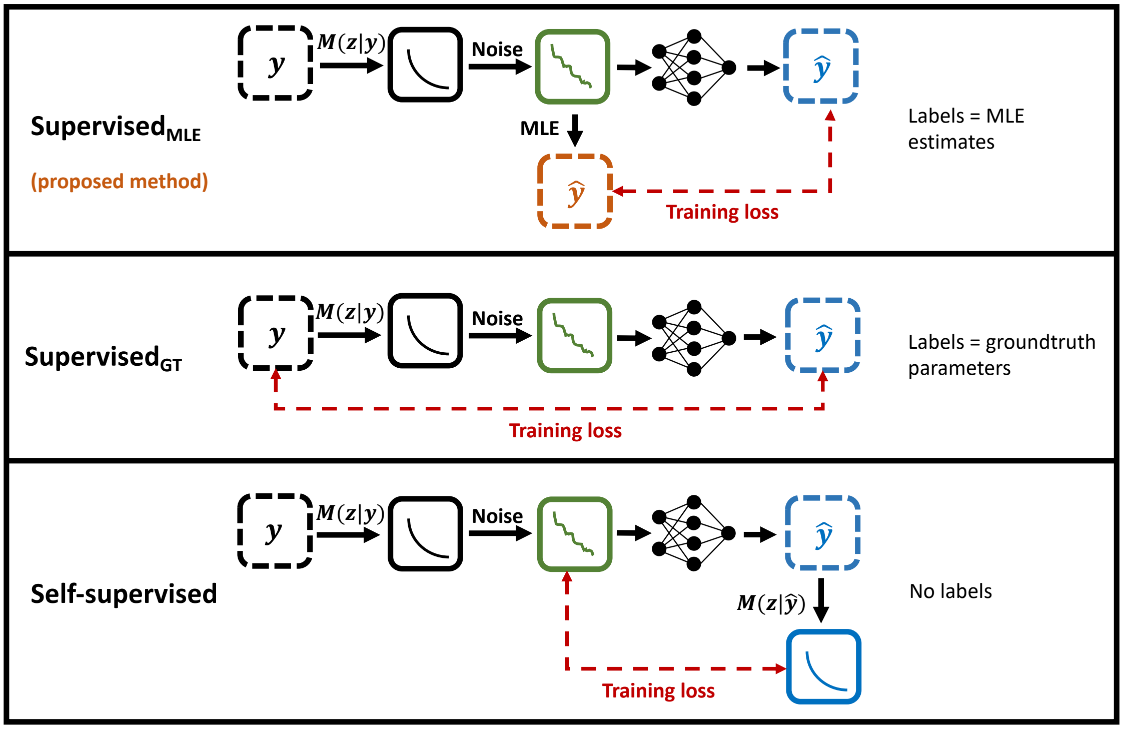

Figure 4-2: Comparison between the proposed method (*Supervised$_{MLE}$*) and existing supervised and self-supervised approaches.

was converted into supervised form by training *Supervised$_{MLE}$* on Gaussian-model based MLE labels. The models are summarised in Table 4.1.

Network name	Loss space	Label	Label noise model
Supervised$_{GT}$	Y	Groundtruth	N/A
Self-supervised	X	N/A	N/A
Supervised$_{MLE,\ Rician}$	Y	MLE	Rician
Supervised$_{MLE,\ Gaussian}$	Y	MLE	Gaussian

Table 4.1: Summary of evaluated parameter estimation networks. Y denotes parameter space; X denotes signal space.

All networks were trained and tested on the same datasets; differences in performance can be attributed solely to differences in loss function formulation and training label selection.

Signal model

As in Chapters 2 and 3, the IVIM model was investigated as an exemplar 4-parameter non-linear qMRI model, which poses a non-trivial model fitting problem has been extensively discussed in the DL qMRI literature [60, 61, 85, 87, 88]:

$$S(b|S_0, f, D_{slow}, D_{fast}) = S_0(fe^{-b(D_{fast}+D_{slow})} + (1-f)e^{-bD_{slow}}) \tag{4.12}$$

where S corresponds to the signal model M, b corresponds to the sampling scheme z, and $[S_0, f, D_{slow}, D_{fast}]$ corresponds to the parameter-vector y. In physical terms, S_0 is an intensity normalisation factor, f is a perfusion fraction, D_{slow} is the diffusivity of non-perfusing water, and D_{fast} is the pseudo-diffusivity of perfusing water.

Network architecture

Network architecture was harmonised across all network variants, and represents common choices made in the qMRI literature [61]: 3 fully connected hidden layers, each with a number of nodes matching the number of signal samples z (i.e. b-values), and an output layer with a number of nodes matching the number of model parameters. Wider (150 nodes per layer) and deeper (10 hidden layers) networks were investigated and found to have equivalent performance, during both training and testing, at the cost of increased training time. All networks were implemented in Pytorch 1.9.0 with exponential linear unit activation (ELU) functions [89]; ELU performance is similar to ReLU, but is more robust to poor network weight initialisation.

Training data

Training datasets were generated at SNR $= [15, 30]$ to investigate parameter estimation performance at both high and low noise levels. At each SNR, 100,000 noise-free signals were generated from uniform IVIM parameter distributions ($S_0 \in [0.8, 1.2]$, $f \in [0.1, 0.5]$, $D_{slow} \in [0.4, 3.0]10^{-3}mm^2/s$, $D_{fast} \in [10, 150]10^{-3}mm^2/s$, representing realistic tissue values), sampling them with a real-world acquisition protocol [73] ($b = [0, 10, 20, 30, 50, 80, 100, 200, 400, 800]$ s/mm^2), and adding Rician noise. Training data generative parameters were drawn from uniform, rather than in-vivo, parameter distributions to minimise bias in network parameter estimation [83]. Data were split 80/20 between training and validation. MLE labels were calculated using a bound-constrained non-linear fitting algorithm, implemented with *scipy.optimise.minimize*, using either Rician log-likelihood or sum of squared errors as fitting objective function. This algorithm was initialised with groundtruth values to improve fitting robustness and avoid local minima. Training/validation samples associated with 'poor' MLE labels (defined as lying on the boundary of the bound-constrained estimation space) were held out during training and ignored during validation.

Network training

Network training was performed using an Adam optimiser (learning rate = 0.001, betas = (0.9, 0.999), weight decay = 0) as follows: $Supervised_{GT}$ (at SNR 30) was trained 16 times on the same data, each time initialising with different network weights, to improve robustness to local minima during training. From this set of trained networks, a single $Supervised_{GT}$ network was selected on the basis of validation loss. The trained weights of this selected network were subsequently used to initialise all other networks; in this way, any differences in network performance could be solely attributed to differences in training label selection and training loss formulation. In the case of supervised loss formulations, the inter-parameter weight vector $\mathcal{W}$ was chosen as the inverse of each parameter's mean value over the training set, to obtain equal loss weighting across all four IVIM parameters.

Testing data and rationale

Networks were tested on both synthetic and real qMRI data. The synthetic approach offers (a) known parameter groundtruths to assess estimation against, (b) arbitrarily large datasets, and (c) tunable data distributions, but is based on possibly simplified qmri signals. This approach was used to assess parameter estimation performance in a controlled, rigorous manner; real data was subsequently used to validate the trends observed in silico.

Unlike synthetic data, in-vivo datasets do not contain 'groundtruth' values against which parameter estimates can be readily assessed. The qMRI literature contains a range of validation strategies, compiled in Appendix E, to address this deficit. This work employs the strategy described in Section E.2.1 which, at the cost of significantly increased acquisition cost, most-closely mimics synthetic testing[3]: extracting reference pseudo-'groundtruths' from super-sampled in-vivo data.

Synthetic data was generated with sampling, parameter distributions, and noise levels matching those used in network training. The IVIM parameter space in which the networks were trained was uniformly sub-divided 10 times in each dimension, to analyse estimation performance as a function of parameter value. At each point in the parameter space, 500 corresponding noisy signals were generated and used to

[3] Excluding ex-vivo histology, discussed in Appendix E, which was not feasible in the context of human pelvic tissue.

test network performance (against generative groundtruth), accounting for variation under noise repetition.

Real data was acquired from the pelvis of a healthy volunteer on a wide-bore 3.0T clinical system (Ingenia, Philips, Amsterdam, Netherlands), 5 slices, 224 x 224 matrix, voxel size = 1.56 x 1.56 x 7mm, TE = 76ms, TR = 516ms, scan time = 44s per 10 b-values listed in section 4.3.1. For the purposes of assessing parameter estimation methods, we obtained gold standard voxelwise IVIM parameter estimates from a supersampled dataset (16-fold repetition of the above acquisition, within a single scanning session, generating 160 b-values, total scan time = 11m44s). Conventional MLE was performed on this supersampled data to produce best-guess 'groundtruth' parameters. During testing, the supersampled dataset was split into 16 distinct 10 b-value acquisitions, each corresponding to a single realistic clinical acquisition. All images were visually confirmed to be free from motion artefacts. The mismatch in parameter distributions between this in-vivo data (highly non-uniform) and the previously-described synthetic data (uniform by construction) limited the scope for validating our in-silico results. To address this, a final synthetic testing dataset was generated from the in-vivo MLE-derived 'groundtruth' parameters, and was used for direct comparison between real and simulated data.

Evaluation metrics

Parameter estimation performance was evaluated using 3 key metrics: (1) mean bias with respect to groundtruth, (2) mean standard deviation under noise repetition, and (3) RMSE with respect to groundtruth. RMSE is the most commonly used metric to evaluate estimation performance [61, 60], but is limited in its ability to disentangle accuracy and precision; to this end, mean bias and standard deviation were used as more specific measures of network performance.

It is important to note that *all* methods were assessed with respect to groundtruth qMRI parameters, *even those trained on MLE labels*. For these methods, the training and validation loss (MLE-based) differed from the reported testing loss (groundtruth-based).

4.3.2 Results

Synthetic data

The relative performance of all previously-discussed parameter estimation methods is summarised in Figures 4-3 and 4-4. These figures show the bias, variance (represented by its square root: standard deviation), and RMSE of parameter estimates with respect to groundtruth values, reported for each model parameter as a function of its value over the synthetic test dataset; each plotted point represents an average over 500 noise instantiations and a marginalisation over all non-visualised parameters. Marginalisation was required for visualisation of a 4-dimensional parameter space; as described below, this representation was confirmed to be representative of the entire, non-marginalised space.

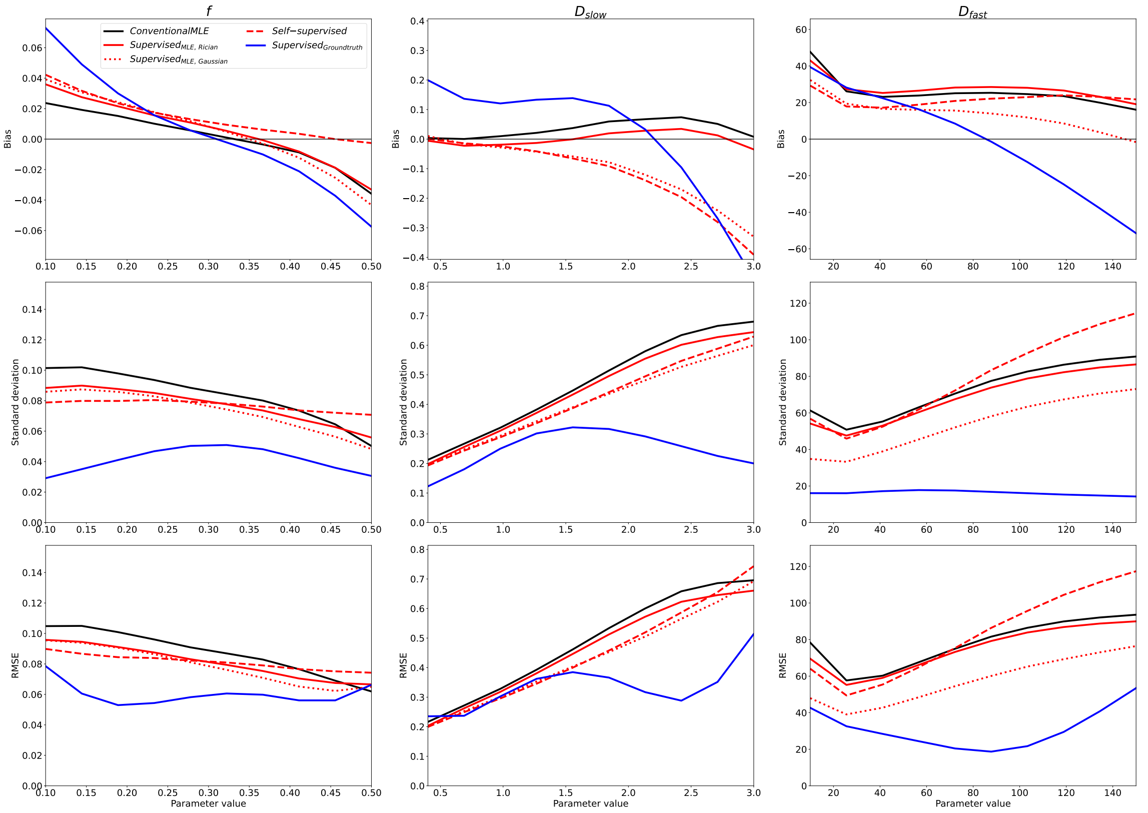

Figure 4-3: Parameter estimation performance at low SNR (15) as a function of groundtruth parameter Y. Performance summarised by bias & RMSE with respect to groundtruth and standard deviation with respect to noise repetition. Conventional MLE fitting is provided as a non-DNN reference standard. For the sake of visualisation, each plotted point represents marginalisation over all non-specified Y dimensions.

In keeping with previously reported results, this work shows a bias/variance trade-off between different parameter estimation methods. Conventional MLE fitting is provided as a reference (plotted in black). Approaches which, on a theoretical level, approximate conventional MLE (*Self-supervised* and $Supervised_{MLE}$, plotted in red), are generally associated with low bias, high variance, and high RMSE, whereas groundtruth-labelled supervised methods (plotted in blue) exhibit lower variance and RMSE at the cost of increased bias.

Increases in bias, if consistent across parameter space Y, do not necessarily reduce

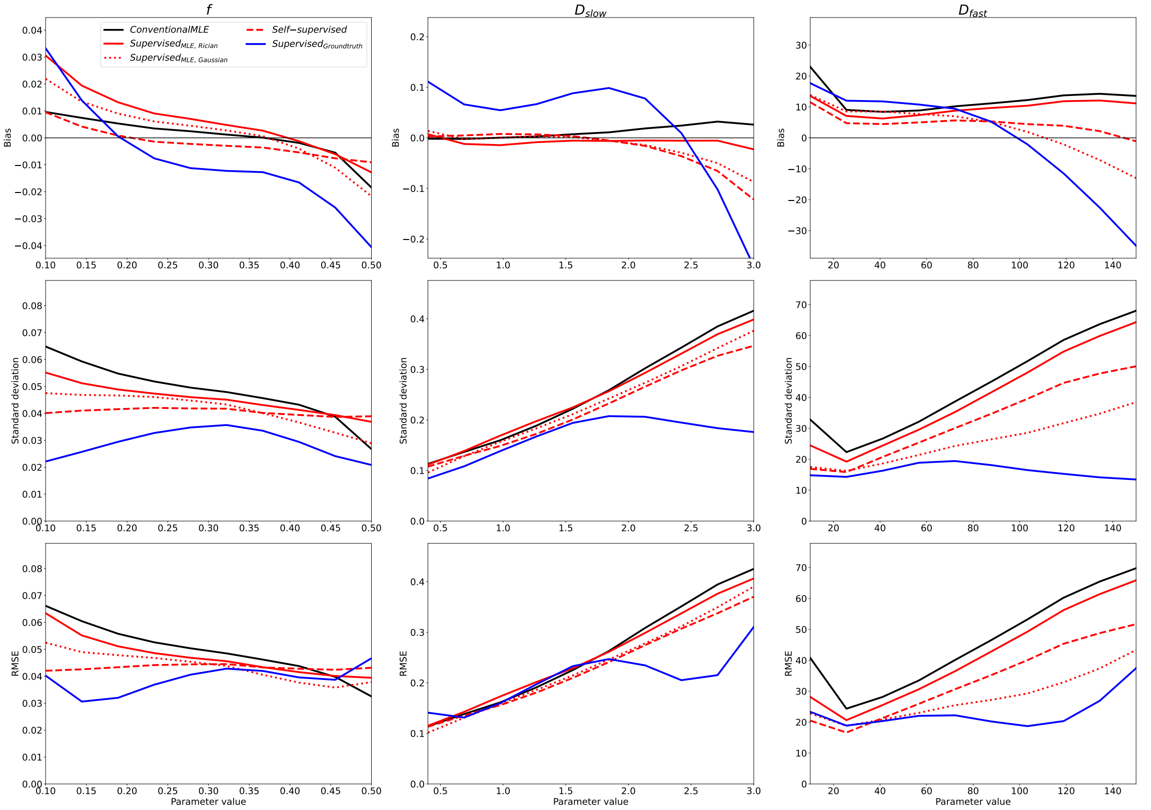

Figure 4-4: Parameter estimation performance, visualised as in Figure 4-3, but for high SNR (30) data.

sensitivity to differences in underlying tissue properties. However, $Supervised_{GT}$ is shown to be associated with bias that *varies significantly* as a function of groundtruth parameter values. This results in a reduction in information content, visualised as the *gradient* of the bias plots (top row) in Fig 4-3. The more negative the gradient, the more parameter estimates are concentrated in the centre of the parameter estimation space $\overline{X}$, and the lower the ability of the method to *distinguish* differences in tissue properties. This information loss can be seen in Fig 4-5, which compares $Supervised_{GT}$ to conventional MLE fitting, and shows the compression in $\overline{X}$ over the groundtruth parameter-space X.

Clinical data

The above trends, found in simulation, were also observed in real-world data. Fig 4-6 shows the bias, variance, and RMSE of parameter estimates with respect to 'groundtruth' values (obtained from the supersampled dataset described in §4.3.1). The x axes of these plots correspond to these reference values. To aid visualisation, 10 uniform bins were constructed along each parameter dimension, into which clinical

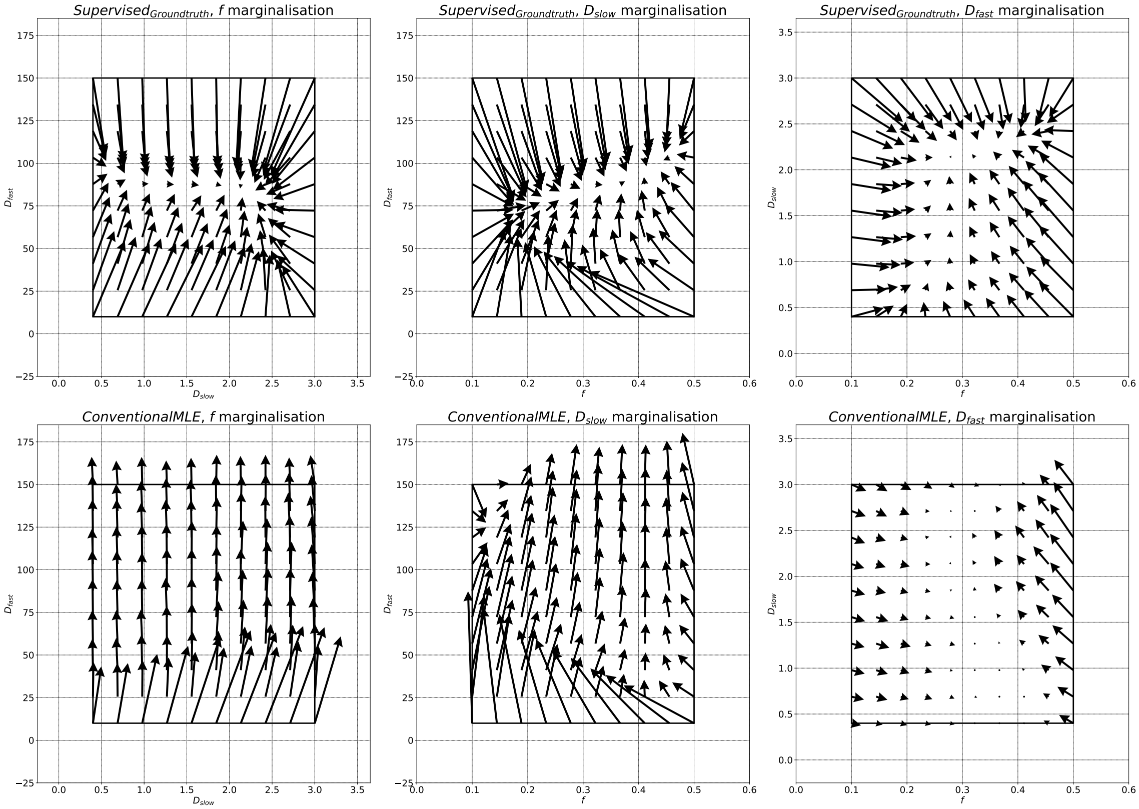

Figure 4-5: Comparison between $Supervised_{GT}$ and reference conventional MLE fitting, expressed in terms of estimation bias and information compression at low SNR (15). Arrows represent the mean mapping from Y to $\overline{Y}$, averaged over noise, as a function of parameter space Y. For the sake of visualisation, each plotted point represents marginalisation over all non-specified Y dimensions.

voxels were assigned based on their 'groundtruth' parameter values. Fig 4-6 plots the mean bias, standard deviation, and RMSE associated with each bin as a function of the bin's central value, together with the distribution of voxels across the 10 bins.

By calculating the variance of the 16 $b = 0$ images, the SNR of this clinical dataset was found to be ∼15; Fig 4-3 is therefore the relevant point of comparison. It can be readily seen that the trends observed in simulated data, described in §4.3.2, are replicated for $f < 0.40$, $D_{slow} < 1.5$, and the entire range of D_{fast}, namely the regions of parameter-space which are well-represented in the real-world data. Fig 4-7 confirms that divergence outside of these ranges is due to under-representation in the in vivo test data; the apparent divergences can be replicated in-silico by matching real-world parameter distributions.

Fig 4-8 contains exemplar parameter maps from the clinical test data, and shows the real-world implications of the trends summarised in Figures 4-3 and 4-6: $Supervised_{GT}$'s low-variance, low-RMSE parameter estimation results in artificially smooth IVIM maps biased towards mean parameter values.

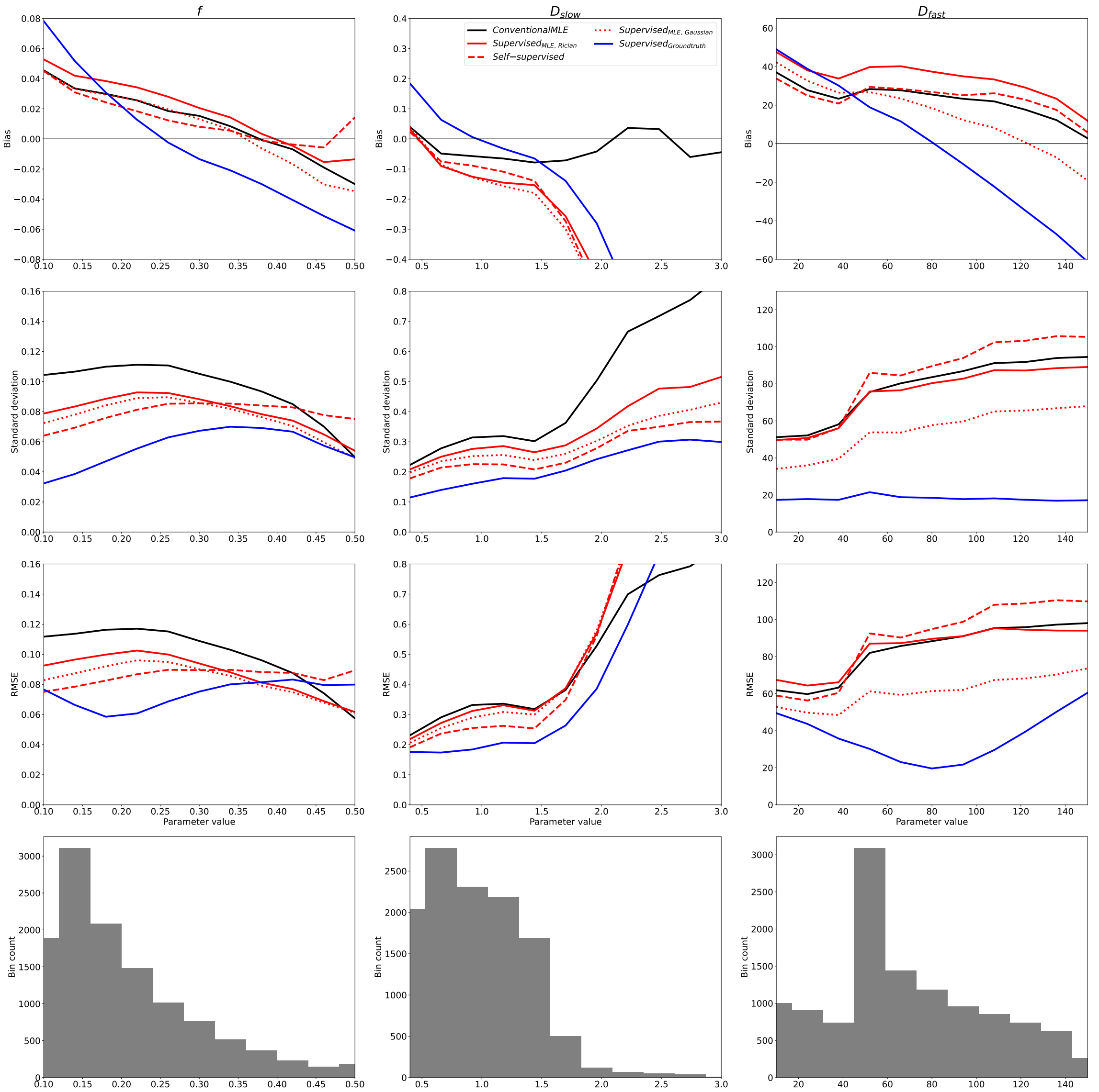

Figure 4-6: In-vivo parameter estimation performance of networks trained on low SNR (15) synthetic data, as a function of supersampling-derived reference parameter values. The first three rows summarise performance by showing bias & RMSE with respect to reference value and standard deviation with respect to noise repetition, marginalised over all non-specified Y dimensions. The bottom row shows the distribution of reference parameter values across the parameter range being visualised.

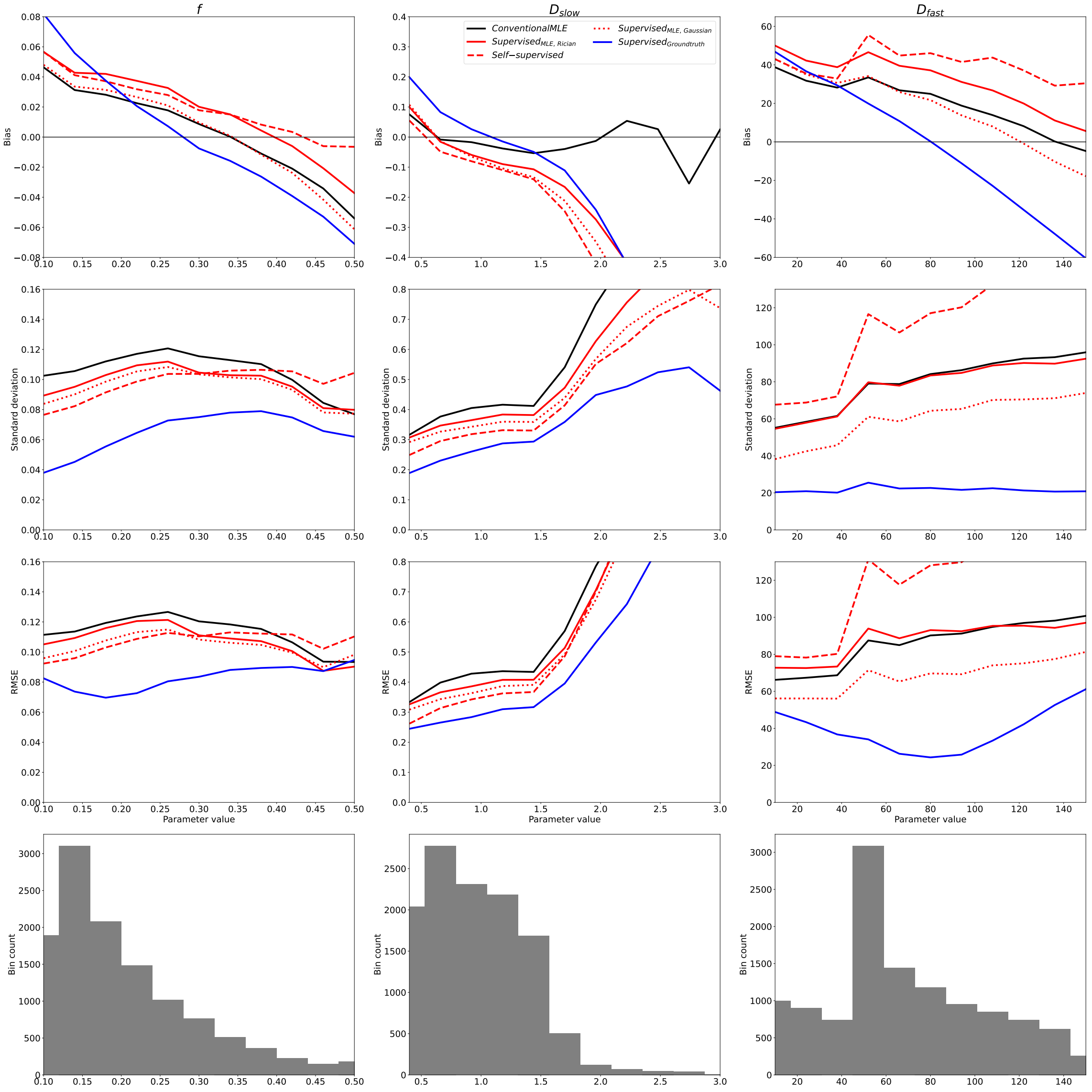

Figure 4-7: Parameter estimation performance of networks trained on low SNR (15) synthetic data, tested on a synthetic dataset matching the distribution of in vivo reference parameter values. The first three rows summarise performance by showing bias & RMSE with respect to groundtruth value and standard deviation with respect to noise repetition, marginalised over all non-specified Y dimensions. The bottom row shows the distribution of groundtruth parameter values across the parameter range, which matches the in vivo dataset by construction.

Advantages offered by the proposed method

The proposed method occupies the low-bias side of the bias-variance trade-off discussed in §4.3.1, and offers four broad advantages over the competing method in this

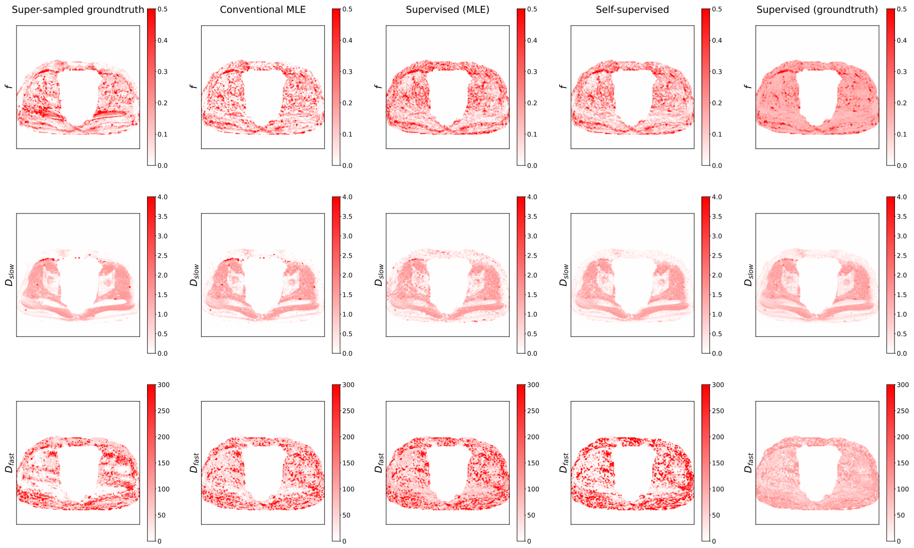

Figure 4-8: Parameter estimation performance of networks on real-world test data, visualised as spatial maps. Groundtruth maps are taken as the maximum likelihood parameter estimates associated with the complete 160 b-value dataset, whereas network predictions are obtained from a single 10 b-value subsample.

space (*Self-supervised*): (i) flexibility in choosing inter-parameter loss weighting $\mathcal{W}$, (ii) incorporation of non-Gaussian (e.g. Rician) noise models, (iii) compatibility with complex, non-differentiable signal models M, and (iv) ability to interface with low-variance methods, to produce a hybrid approach tunable to the needs of the task at hand. These advantages are analysed in turn.

(i) Choice of inter-parameter loss weighting $\mathcal{W}$

By computing loss in parameter-space Y, the proposed method has total flexibility in adjusting the relative contribution of different y in the training loss function. In contrast, since *Self-supervised* calculates training loss in X, the relative weighting depends on the acquisition protocol z. Fig 4-9 compares the proposed method - weighted so as to not discriminate between different model parameters - with variants designed to overweight single parameters by a factor of 10^6. The potential advantages offered by this selective weighting are seen in the estimation D_{fast}, where this approach leads to a small increase in both precision and accuracy. This parameter-specific weighting is not available within a *Self-supervised* framework.

In light of the differences arising from inter-parameter loss weighting, subsequent analysis uses $Supervised_{MLE,\ Gaussian}$ as a proxy for *Self-supervised*; both methods encode the same regularised MLE fitting, but differ in their inter-parameter weighting.

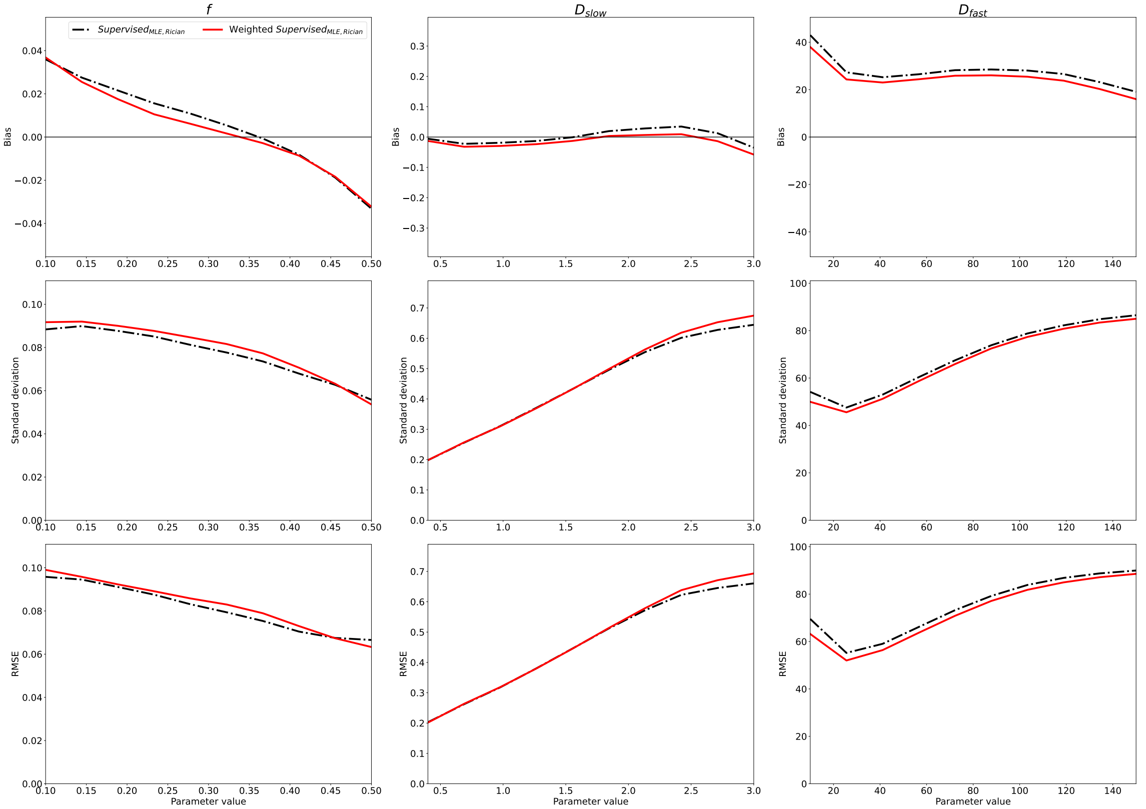

Figure 4-9: Comparison between $Supervised_{MLE,\ Rician}$, as described above, and variants which differ in their inter-parameter loss weighting $\mathcal{W}$, at low SNR (15). Each column compares $Supervised_{MLE,\ Rician}$ to a different network variant, uniquely trained to overweight the single relevant signal model parameter. For the sake of visualisation, each plotted point represents marginalisation over all non-specified Y dimensions.

(ii) Incorporation of Rician noise modelling

By pre-computing MLE labels using conventional parameter estimation methods, the proposed method is able to incorporate accurate Rician noise modelling. Comparison between $Supervised_{MLE,\ Rician}$ and $Supervised_{MLE,\ Gaussian}$ shows the effect of the choice of noise model; these differences are most pronounced at low SNR (Fig 4-3) and high D_{slow}, when the Gaussian approximation of Rician noise is known to break down. In this regime, the proposed method gives less biased, more informative D_{slow} estimates, replicating conventional MLE performance at a fraction of the computational cost. At high D_{slow}, it has a flatter, more information-rich, D_{slow} bias curve than all other DL methods. The information loss associated with incorrect noise modelling is further visualised in Fig 4-10, which shows the compression in D_{slow} estimates $\overline{X}$ over the groundtruth parameter-space X. As expected, this compression is most apparent at high values of D_{slow}, when the signal is more likely to approach the Rician noise floor.

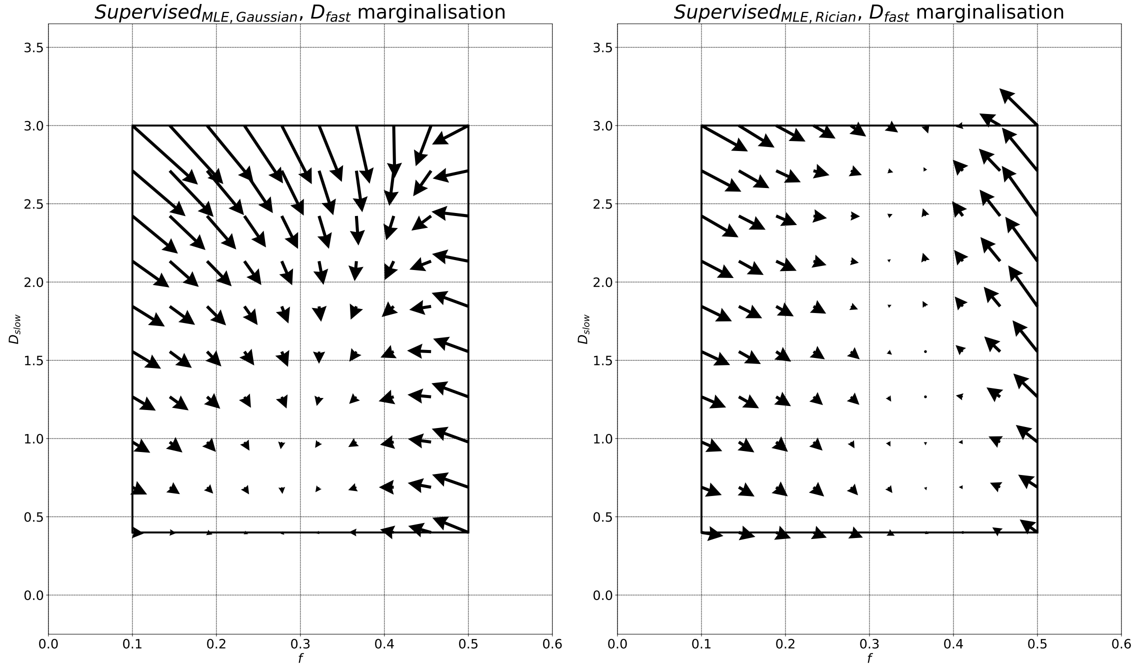

Figure 4-10: Comparison of the information content captured by $Supervised_{MLE}$ methods, as a function of the noise model used in computing MLE labels, at low SNR (15). Arrows represent the mean mapping from Y to $\overline{Y}$, averaged over noise, as a function of parameter space Y. For the sake of visualisation, each plotted point represents marginalisation over all non-specified Y dimensions.

(iii) Compatibility with complex signal models

An additional advantage of computing training loss in parameter-space Y is that DNN networks are signal model agnostic: network training does not require explicit calculation of M. This approach is advantageous when working with complex signal models, as made clear by comparison with *Self-supervised* methods. In contrast with the proposed approach, *Self-supervised* methods embed M between network output and training loss (see Fig 4-2). During DNN training, network parameters p are updated by computing partial derivatives of the training loss; this process requires the loss to be expressed in a differentiable form. By embedding M in the loss formulation, *Self-supervised* methods thus limit themselves to signal models that can be expressed in an explicitly differentiable form. In contrast, the proposed method does not rely on computing M during training, and is therefore compatible with a wider range of complex qMRI signal models.

(iv) Tunable network approach

As discussed above, this work shows a clear bias/variance trade-off between different parameter estimation methods. The optimal choice of method depends on the task at hand [90], and may not lie at either extreme of this trade-off. Therefore, it would be advantageous to be able to combine low-bias and low-variance methods into a single, hybrid approach, with performance tunable by the relative contribution of each constituent method. The proposed method, which interfaces naturally with $Supervised_{GT}$, offers exactly that. An example of this approach is shown in Fig 4-11: training loss has been weighted equally between groundtruth and MLE labels, and, as expected, the resulting network performance lies in the middle ground between these two extremes.

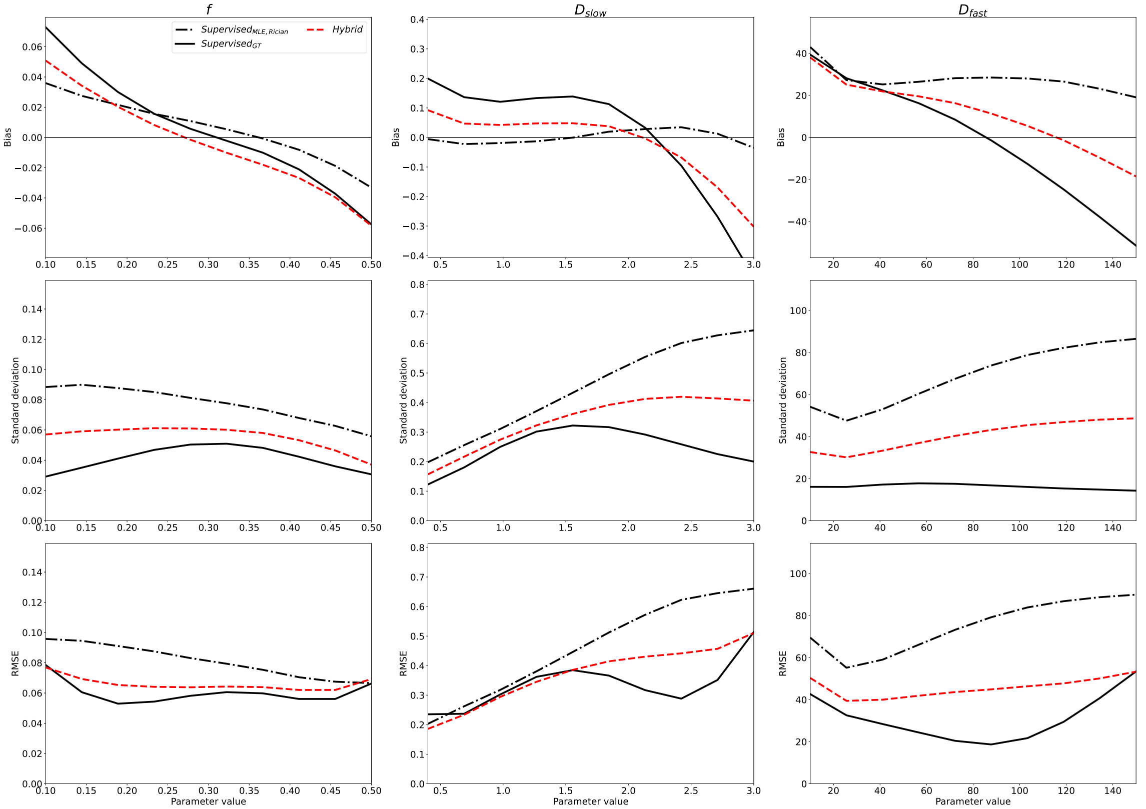

Figure 4-11: Proof of concept of a hybrid parameter estimation method, formed by training a supervised network with an equally-weighted sum of $Supervised_{MLE,\ Rician}$ and $Supervised_{GT}$ loss functions, at low SNR (15). For the sake of visualisation, each plotted point represents marginalisation over all non-specified Y dimensions.

Comparison with conventional fitting

Comparison between the proposed method ($Supervised_{MLE,\ Rician}$) and conventional fitting (*MLE, Rician*) highlights additional advantages offered by our approach. Firstly, Figs 4-3 and 4-4 demonstrate qualitatively similar performance between these methods across the entire parameter space. The fact that the proposed method, which offers near-instantaneous parameter estimation, produces similar parameter estimates to well-understood conventional MLE methods justifies its adoption in and of itself. However, it not only mimics but indeed in many cases outperforms (lower bias, variance, and RMSE) the very same method used to compute those labels. This finding not only motivates its use, but also confirms that DL methods are able to exploit information shared between training samples beyond what would be possible by considering each sample in isolation.

A note on RMSE

RMSE is a poor summary measure of network performance. It is heavily skewed by outliers, and thus favours methods which give parameter estimates consistently close to mean parameter values. Such estimates, as in the case of D_{fast}, may contain very little information (Fig 4-5) despite being associated with low RMSE. Accordingly, this work strongly recommends that RMSE be discontinued as a single summary metric for parameter estimation performance: it must always be accompanied by bias, variance, and ideally an analysis of information content.

Justification of parameter marginalisation

The above analysis has been largely based on Figs 4-3 and 4-4, which show parameter estimation performance marginalised over 3 dimensions of X. This choice, made to aid visualisation, was validated against higher dimensional representations of the same data.

Fig 4-12 compares $Supervised_{MLE,\ Rician}$ and $Supervised_{Groundtruth}$ performance across the entire qMRI parameter space. It can be seen that trends observed in Fig 4-3 are replicated here; attention is drawn to two such examples. Firstly, Fig 4-3 suggests that $Supervised_{Groundtruth}$ produces lower f standard deviation than $Supervised_{MLE,\ Rician}$; Fig 4-12 confirms this to be the case across all test data. In contrast, Fig 4-3 suggests that $Supervised_{Groundtruth}$ produces higher D_{slow} bias at low D_{slow} and lower bias at high D_{slow}; Fig 4-12 confirms a spread of bias differences across the test data: some favouring one method, and others the other. This effect is explored in Fig 4-13, which compares D_{slow} estimation performance as a function of f and D_{fast} at two specific (non-marginalised) groundtruth D_{slow} values $(0.69, 2.71)$. As expected from the marginalised representation in Fig 4-3, at low D_{slow} $Supervised_{Groundtruth}$ produces higher bias *across the entire* f-D_{fast} *parameter space*, whereas at high D_{slow} the opposite is true.

Despite this, it is important to note the limitations of marginalisation. Fig 4-13 also shows that the relative performance of $Supervised_{MLE,\ Rician}$ and $Supervised_{Groundtruth}$ varies across *all parameter-space dimensions*. Consider $D_{slow} = 0.69$, where Fig 4-3 shows similar marginalised RMSE for these methods. In fact, by visualising this difference as a function of f and D_{fast}, we reveal two distinct regions: high f/low D_{fast} (where $Supervised_{MLE,\ Rician}$ produces lower RMSE), and elsewhere (where it produces higher RMSE). This highlights (i) the potential pitfalls of producing summary results by marginalising across entire parameter spaces and (ii) the need to choose parameter-estimation methods appropriate for the specific parameter combinations relevant to the tissues being investigated [90].

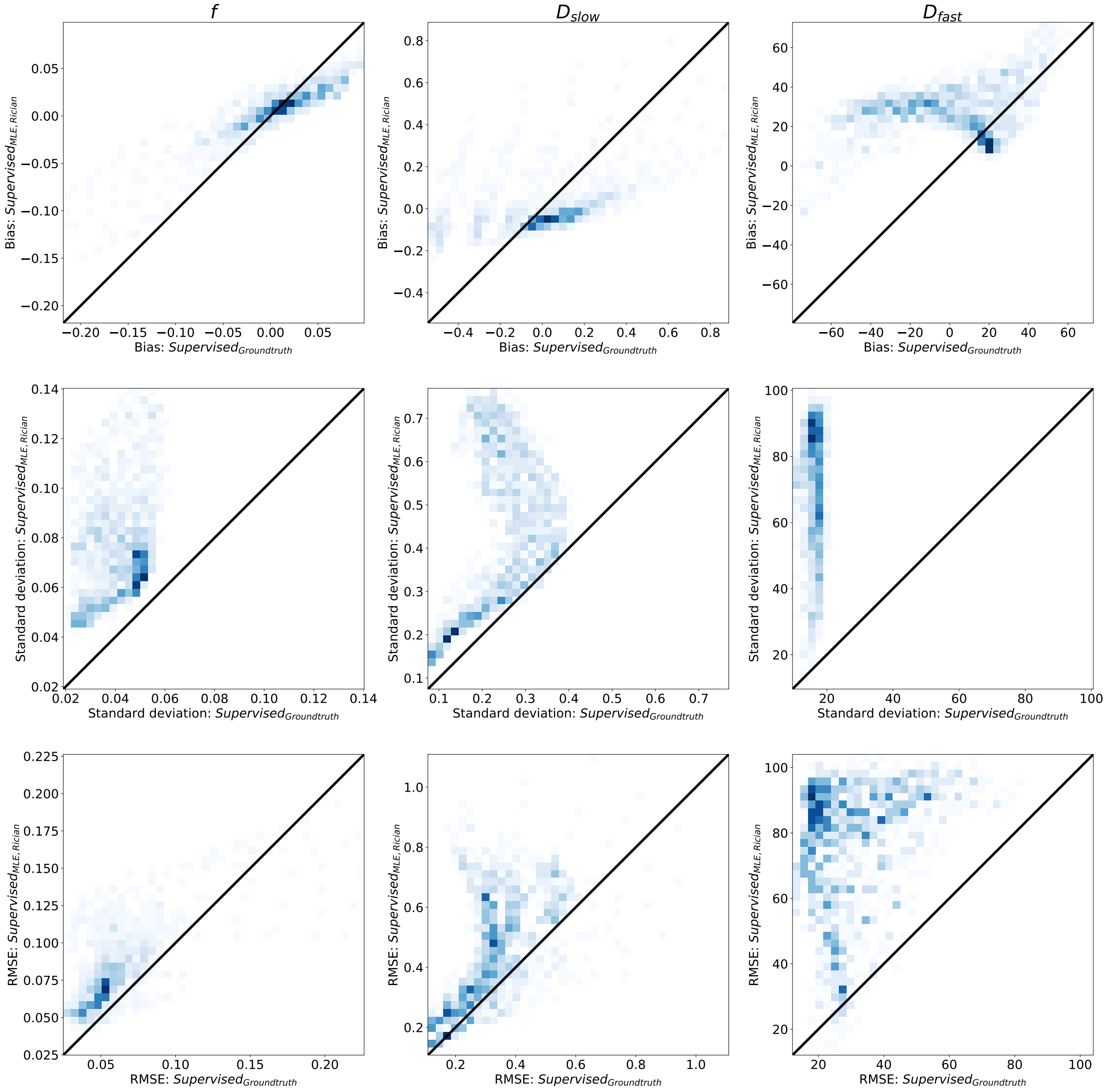

Figure 4-12: Non-marginalised comparison of parameter estimation performance between $Supervised_{MLE,\ Rician}$ and $Supervised_{Groundtruth}$ at low SNR (15). Colour intensity represents density of distribution across all X and all noise repetitions.

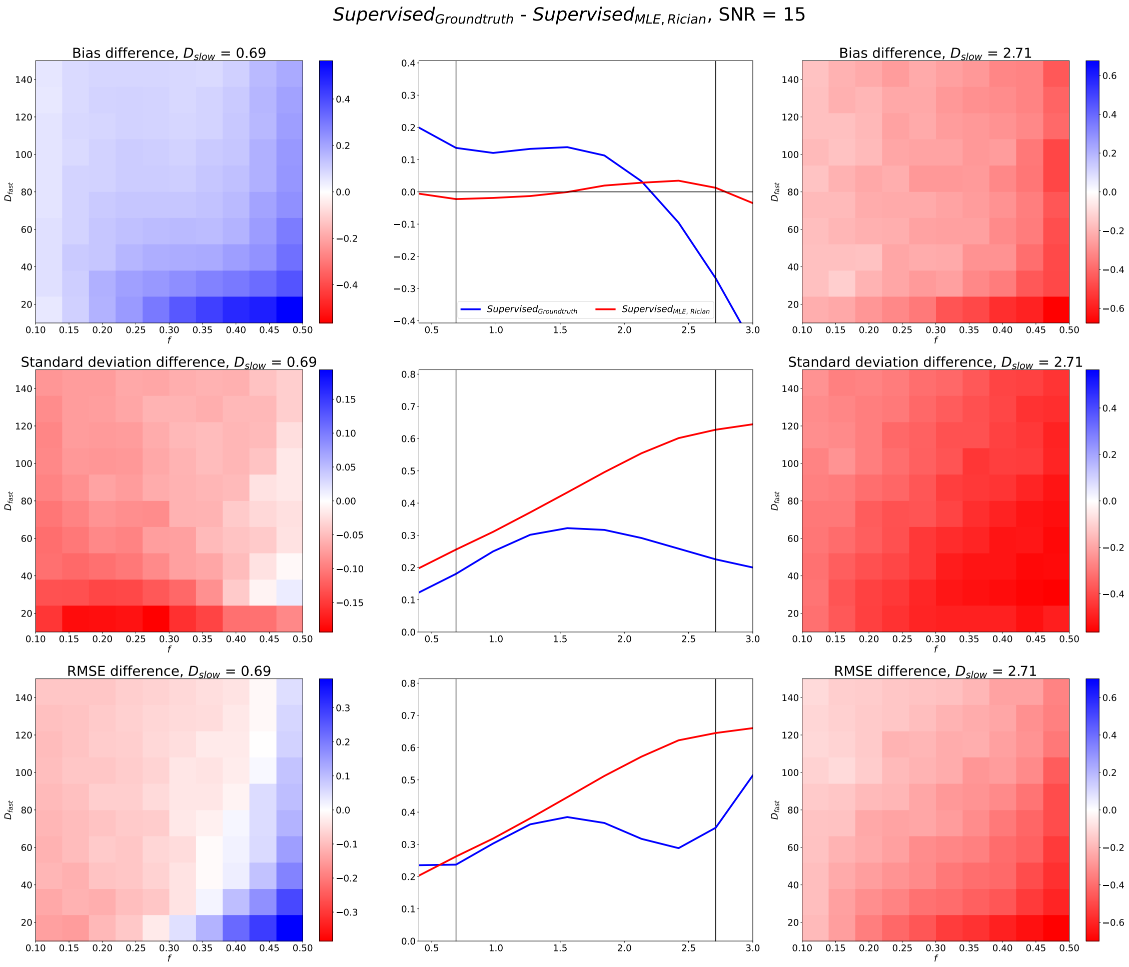

Figure 4-13: Differences in performance (bias, standard deviation, RMSE) between $Supervised_{MLE,\ Rician}$ and $Supervised_{Groundtruth}$ for two groundtruth values of D_{slow} at low SNR (15). The outermost columns (left and right) correspond to $D_{slow} = 0.69$ and $D_{slow} = 2.71$ respectively, and show mean performance under noise repetition, without marginalisation. The central column reproduces the corresponding marginalised representation from Fig 4-3.

4.4 Discussion and conclusions

This work draws inspiration from state-of-the-art supervised and self-supervised qMRI parameter estimation methods to propose a novel DNN approach which combines their respective strengths. In keeping with previous work, a bias/variance trade-off is observed between existing methods; supervised training produces low variance under noise, whereas self-supervised leads to low bias with respect to groundtruth.

The increased bias of supervised DNNs is counter-intuitive: when labels are available, these methods have access to more information, and should therefore outperform, non-labelled alternatives. In light of this, this work proposes that the high bias associated with these supervised methods stems from the *nature* of the additional information they receive: groundtruth training labels. By careful adjustment of these labels, this work shows that the low-bias performance previously limited to self-supervised approaches can be achieved within a supervised learning framework.

This framework forms the basis of a novel low-bias supervised learning approach to qMRI parameter estimation: training on conventionally-derived maximum likelihood parameter estimates[4]. This method offers four clear advantages to competing non-supervised low-bias DNN approaches: (i) flexibility in choosing inter-parameter loss weighting, which enables network performance to be boosted for qMRI parameters of interest; (ii) incorporation of Rician noise modelling, which improves parameter estimation at low SNR; (iii) separation between signal model and training loss, which enables the estimation of non-differentiable qMRI signal models; and, crucially, (iv) ability to interface with existing supervised low-variance approaches, to produce a tunable hybrid parameter estimation method.

4.4.1 Implications & outlook

This final point - interfacing with complementary groundtruth-labelled methods - constitutes the key contribution of this work: unifying low-bias and low-variance parameter estimation under a single supervised learning umbrella. When faced with a parameter estimation problem, we no longer need to choose between extremes of the bias/variance trade-off; rather, we can tune DNN parameter estimation performance to the specific needs of the task at hand.

This hybrid approach is a perfect match for the task-driven qMRI CED paradigm presented in Chapters 2 and 3; θ_{est} can for the first time be adjusted smoothly on a task-by-task basis.

4.4.2 Next steps

Combined, the contributions presented in the thesis thus far give us the basic tools needed to realise the vision of task-driven qMRI computational experimental design.

[4] This work has focused on voxelwise DL parameter estimation methods: networks which map one signal curve to its corresponding parameter estimate. There are, however, alternatives: convolutional neural network methods which map spatially-related clusters of qMRI signals to corresponding clusters of parameter estimates [91, 92, 93]. The proposed MLE training label approach could be incorporated into such methods, and it is left to future work to investigate the effect this would have on parameter estimation performance.

In Chapter 5, the novel parameter estimation method introduced above is combined with the task driven pipeline proposed in Chapter 2 and validated in Chapter 3. Conventional parameter estimators are assessed against their ML counterparts on the metric that truly matters: task performance. This demonstration of θ_{est} assessment, comparison, and selection can serve as a blueprint for those convinced by the argument that, as a qMRI field, we should implement *task-driven* CED.

Chapter 5

A blueprint for task-driven experimental design: choosing experimental settings

Contents

This chapter is adapted from:

- **Epstein SC**, Bray TJP, Hall-Craggs M and Zhang H, *Do deep learning-based qMRI parameter estimators improve clinical task performance?*, 2023, ISMRM Annual Meeting 2023

This thesis has (i) made the case for task-driven qMRI CED and (ii) provided two key tools necessary for its implementation: a holistic computational pipeline, described in Chapters 2 and 3; and a low-bias parameter estimator, described in Chapter 4. This chapter combines these tools and concepts, and provides a blueprint for performing task-driven computational experimental design in the real world.

5.1 A blueprint for task-driven qMRI CED

qMRI experiments are defined by three experimental settings: acquisition protocol (θ_{acq}), signal model (θ_{mod}), and parameter estimation method (θ_{est}). Appropriate selection of these settings leads to improved experimental utility ('task performance'). Selection requires assessment, which, in light of the settings' *interactions with each other*, should be performed *holistically*.

This assessment/selection process involves:

1. Identifying a qMRI *task* to be tackled, such as disease classification or tissue property measurement.
2. Choosing a quantitative *quality metric* by which this task's performance can be assessed and optimised.
3. Defining a *population of interest* which describes the tissues that one expects to encounter when tackling the task; the experimental choices are optimised for this population.
4. Proposing a range of *plausible experimental settings* (θ_{acq}, θ_{mod}, θ_{est}) which, when combined, produce candidate experimental designs.
5. Predicting the *task performance* associated with each of these candidate experimental designs, using the pipeline described in Chapter 2.
6. Selecting the experimental design associated with the *largest task performance*.

5.2 Two worked examples: selecting parameter estimators

What follows are two worked examples (labelled 'W1' and 'W2') of the CED process described above. For simplicity and clarity, their scope is limited to selecting a parameter estimation method (θ_{est}) in the context of pre-selected signal model and acquisition protocol.

W1 and W2 serve not only as exemplars of task-driven CED, but also represent the first time DL-based qMRI parameter estimation methods have been assessed against their conventional counterparts on their ability to solve real-world clinical tasks.

5.2.1 Identifying a qMRI task

The qMRI tasks to be tackled involve classifying SpA lesions using dMRI; W1 and W2 correspond to E1 and E2 described in Section 2.4.2. W1 involves classifying suspected SpA lesions as either 'healthy' or 'chronic', and W2 consists of classifying lesions as 'active' or 'chronic'.

This classification is performed by acquiring diffusion-weighted data, fitting the IVIM dMRI model to that data, and classifying tissues based on their best-fit parameter estimates: the perfusion fraction f in the case of W1, and the diffusivity D_{slow} in the case of W2.

5.2.2 Choosing a quality metric

The AUC of the ROC curve associated with the chosen IVIM parameter is selected as the task-performance quality metric for both tasks.

5.2.3 Defining a population of interest

The generative IVIM parameters associated with the tissue populations of interest are listed in Table 5.1.

Task	**Tissue**	**Signal model**	$\boldsymbol{f}$	$\boldsymbol{D_{slow}}$ $(10^{-3} mm^2/s)$	$\boldsymbol{D_{fast}}$ $(10^{-3} mm^2/s)$	**SNR**	**Sampling** (s/mm^2)	**Parameter of interest**	**Quality metric**
W1	Healthy	IVIM	0.09	0.35	123	20	0, 10, 20, 40, 80, 100, 200, 400, 600	f	AUC
	Chronic		0.12	0.35	123				
W2	Active		0.12	0.60	123				
	Chronic		0.12	0.46	123				

Table 5.1: Experimental settings associated with the worked examples W1 and W2.

5.2.4 Proposing plausible experimental settings

Acquisition protocol and signal model are fixed as detailed in Table 5.1. For both W1 and W2, five parameter estimation methods (θ_{est}) are combined with these predetermined settings:

1. $\boldsymbol{Supervised_{MLE}}$
   A supervised DL parameter estimation method trained on MLE-derived labels.

2. $\boldsymbol{Supervised_{GT}}$
   A supervised DL parameter estimation method trained on groundtruth generative labels.

3. ***Self-supervised***
   A self-supervised DL parameter estimation method trained without explicit labels.

4. ***bcNLLS***
   An iterative MLE-like method which places explicit bound-constraints on the estimated model parameters.

5. ***sNLLS***
   A regularised two-step 'segmented' MLE-like method specific to the IVIM model. The first step fits an ADC model to data above a *threshold diffusion weighting*. This high-diffusion-weighted data retains signal from only the 'slow', non-perfusing IVIM water compartment[1]. The IVIM parameter estimate D_{slow} is subsequently fixed to this ADC value, and bcNLLS is performed on the three remaining IVIM parameters.

Detailed descriptions of these methods are found in Chapters 2 and 4.

5.2.5 Computing task performance

Following the pipeline outlined in Chapter 2, 25,000 noisy IVIM signals were synthesised for each task's SpA subtypes. The IVIM model was fit to these signals using the 5 aforementioned parameter estimation methods.

The resulting IVIM parameter estimates were used to calculate ROC curves for the classification task. Each method's task performance was summarised by the ROC-derived AUC metric.

5.2.6 Selecting the best-performing experimental design

The simulated ROC curves and associated AUC values are shown in Figure 5-1. In both tasks, the parameter estimator associated with the highest task performance is $Supervised_{GT}$. This method is selected for experimental use.

5.3 Looking beyond a summary quality metric

The worked examples above demonstrate the most basic implementation of task-driven CED: identifying the experimental settings that maximise task performance.

[1] See Appendix C for more details.

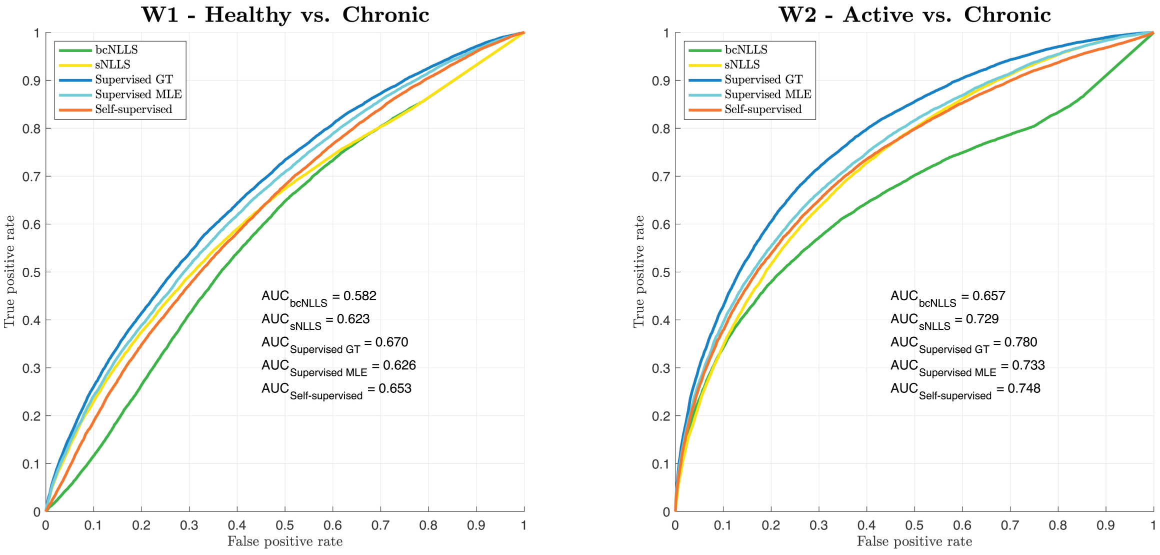

Figure 5-1: ROC curves and associated AUC values for W1 and W2.

However, the simulation framework introduced in Chapter 2 additionally gives insight into *why* certain experimental settings outperform others.

Figure 5-2 shows the noise-induced distribution of parameter estimates on which Figure 5-1 is based. W1 and W2, being classification tasks, rely on being able to distinguish the two distributions represented by solid and dashed lines. In both tasks, conventional estimators (bcNLLS and sNLLS) are associated with broader (higher variance) distributions than DL methods.

The analysis presented in Chapter 4 predicts two classes of DL method: (i) $Supervised_{MLE}$ and $Self\text{-}supervised$ (higher variance, lower absolute and relative bias), and (ii) $Supervised_{GT}$ (lower variance, higher absolute and relative bias); Figure 5-2 confirms this trend. In these tasks, $Supervised_{GT}$ achieves the highest task performance by minimising variance under noise (σ), despite below-average inter-distribution mean separation ($\Delta\mu$).

The fourth step of CED - proposing plausible experimental designs - is crucial for obtaining high-quality qMRI experiments yet prone to user error. The task-driven pipeline, as presented in Chapter 2, can only assess the experimental settings *it is provided with*. If users input a selection of poor experimental settings, the pipeline will necessarily output a ‘bad’ experimental design. Insight into *why* certain experimental settings succeed - or, more importantly, fail - can guide the selection of *good* candidate experimental designs.

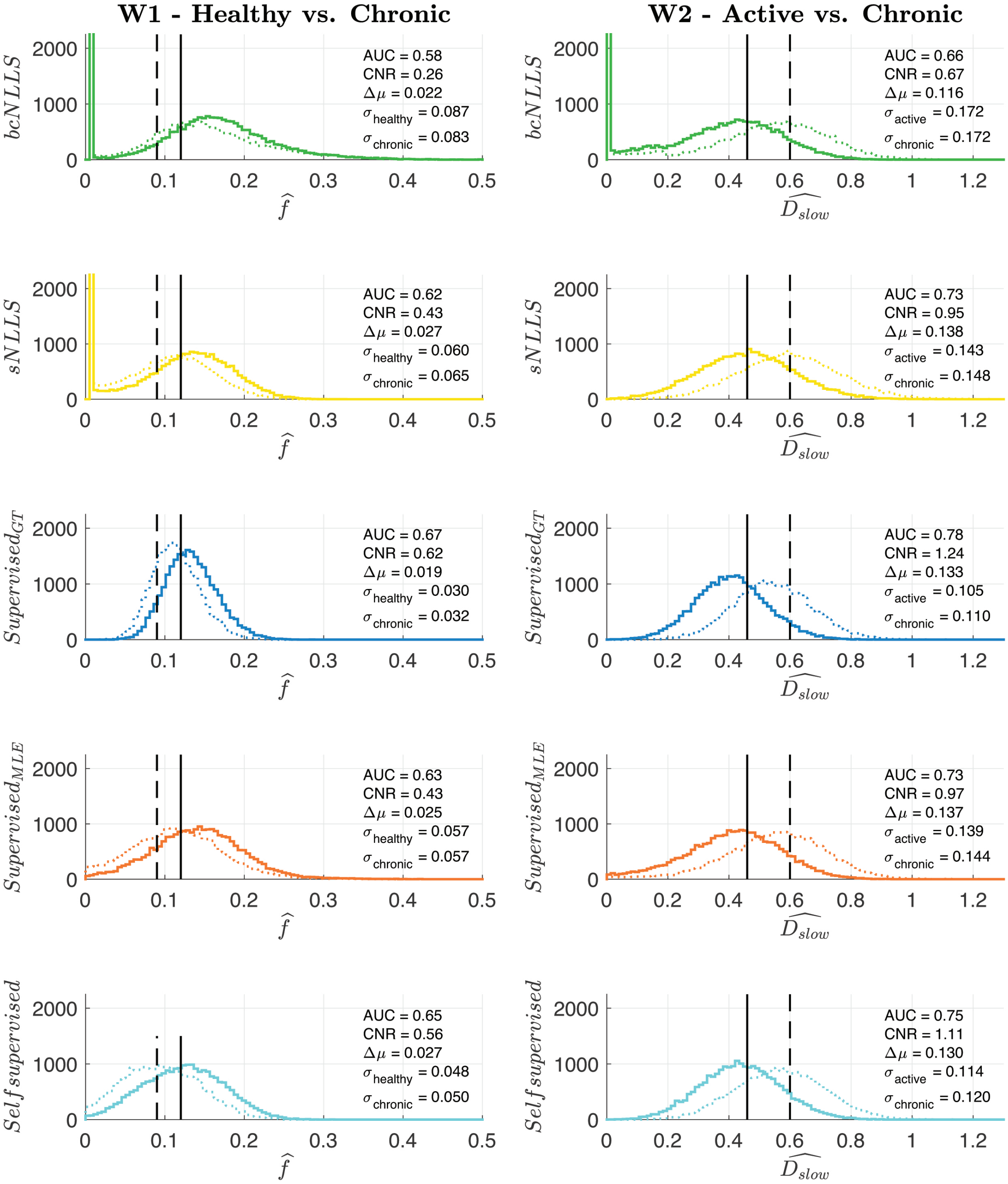

Figure 5-2: Distribution of parameter estimates underpinning the ROC curves shown in Figure 5-1.

5.4 What next?

This chapter has combined the tools introduced in Chapters 2-4 to demonstrate two real-world worked examples of task-driven qMRI CED. This not only provides a blueprint for task-driven qMRI CED but also evaluates, for the first time, DL parameter estimators' classification performance and validates the trends observed in Chapter 4.

Chapter 6 goes one step further, by not only highlighting the key theoretical and practical contributions presented throughout this thesis, but also reflecting on how they may be further refined and expanded upon, cementing *task-driven* optimisation as the default qMRI CED paradigm.

Chapter 6

Conclusions and outlook

Contents

6.1 Thesis summary

This thesis contributes to the field of quantitative magnetic resonance imaging (qMRI) computational experimental design (CED). qMRI CED assesses qMRI experimental designs *in-silico*, enabling the selection of 'good' qMRI experiments without costly in-vivo data acquisition.

This work (i) identifies two broad limitations with the qMRI CED status quo and (ii) presents four contributions to address these shortcomings.

6.1.1 qMRI CED limitations

The first limitation relates to the *quality metrics* used to assess qMRI experimental designs. Existing metrics evaluate qMRI experimental outputs (parameter estimates) on their similarity to tissue-property-derived ‘groundtruths’. This approach is motivated by viewing qMRI as simply ‘tissue measurement’, but does not match common clinical use-cases: performing *tasks*, such as differentiating between tissues in different states of pathology. The ‘quality’, or usefulness, of the parameter estimates arising from such task-based experiments depends on effect size rather than numerical accuracy; task-agnostic groundtruth-based metrics are fundamentally inappropriate.

The second limitation relates to the piecemeal, sequential nature by which CED is performed. qMRI experiments combine three experimental settings: acquisition protocol, signal model, and parameter estimation method. As things stands, CED determines these settings *sequentially* without accounting for how they interact *with each other*. The overall quality of a qMRI experiment - whether that be task performance or tissue measurement - depends on these interactions, and its assessment and optimisation is limited by evaluating experimental settings in isolation.

6.1.2 Thesis contributions

Chapters 2-5 outline four distinct contributions aimed at addressing these shortcomings.

Reimagining the problem: holistic and task-driven

Chapter 2 presents a theoretical analysis of CED, introducing the notion of P_{target}, the parameter estimates produced by an ‘ideal’ qMRI experiment. These are used to re-frame CED’s purpose: from tissue measurement fidelity to task performance maximisation. The case is made that setting $P_{target} = T$, as is standard in current CED practice, is not appropriate for all qMRI experiments; P_{target} should instead be *adjusted* on an experiment-by-experiment basis to reflect the qMRI ‘task’ being performed. Furthermore, it is argued that assessing an experiment’s parameter estimates (i.e. determining how close they are to P_{target}) should explicitly account for the interactions between all components of an experiment. A computational tool which addresses both shortcomings is proposed and implemented: a pipeline which (i) enables P_{target} to be adjusted *task-specifically* and (ii) evaluates experiments’ ability to generate close-to-P_{target} estimates *holistically*. Two exemplar qMRI experiments are used to demonstrate the value of this approach by contrasting it to traditional task-agnostic CED methods.

Validating the method: in-silico vs. in-vivo

Chapter 3 validates the aforementioned pipeline's task performance predictions on in-vivo data. Four qMRI experimental tasks, spanning two clinical datasets, are identified and replicated in-silico. The pipeline's predictions are shown to match - both qualitatively and quantitatively - the 'real-world' task performance. This result, which confirms that the pipeline is able to evaluate experimental designs without needing to acquire costly in-vivo data, has three practical consequences. Firstly, that the pipeline could be used to select the *best* experimental design from a shortlist of candidate experimental settings. Secondly, it motivates the pipeline's further development into an optimisation framework, where experimental settings are iteratively adjusted to maximise task performance. Finally, it validates the demonstration, found in Chapter 2, that task-driven CED approaches outperform task-agnostic ones.

Chapters 2 and 3, between them, provide a framework which address both CED qMRI limitations: task-driven assessment metrics are introduced, and a reliable holistic way of evaluating them is presented. The remainder of the thesis builds on this framework, improving it and demonstrating how it may be applied in the real world.

Improving parameter estimation: a novel deep learning method

The pipeline introduced in Chapter 2 and validated in Chapter 3 can only generate experiments as good as the candidate settings given as input to it. Limitations with one of these settings - choice of parameter estimation method - are addressed in Chapter 4, by introducing a novel DNN-based parameter estimator. This method is inspired by a theoretical analysis of existing parameter estimation methods, which highlights the bias-variance trade-off that underpins existing parameter estimators. This trade-off motivates the proposed method's formulation: a supervised DNN trained on *non-groundtruth* independently-computed MLE labels. This estimator, when evaluated on both in-silico and in-vivo data, is shown to occupy the 'low-bias' side of the aforementioned trade-off, without the compromises associated with competing methods in this space. The Chapter concludes by demonstrating a proof-of-concept application of this method to a hybrid, tunable, and task-driven estimator, designed to unlock the full potential of the pipeline introduced in Chapter 2.

A blueprint for task-driven experimental design: choosing experimental settings

Chapter 5 contains the final two contributions of the thesis. The first is the combination and distillation of the tools developed throughout the preceding Chapters into an end-to-end implementation of task-driven CED. This provides a generalisable blueprint applicable to any qMRI experiment for which the relationship between tissue properties and qMRI biomarkers is known. The second contribution is the first demonstration of the suitability of DL estimators for clinical classification tasks, underpinned by analysis which validates the bias-variance trends described in Chapter 4.

6.2 Outlook

The central message of this thesis, and its most fundamental contribution, is contained in the first half of Chapter 2: a theoretical re-framing of qMRI, away from *measuring* tissue properties and towards a more general notion of *implementing* ‘tasks’. The work which follows, presented in Chapters 3-5, represents only the beginning of what could be achieved with task-driven CED. Outlined below are some areas in which the work presented in this thesis could be expanded to exploit this exciting, new, task-centric paradigm.

6.2.1 Assessment vs. optimisation

The pipeline presented in Chapter 2 and validated in Chapter 3 performs one function: provided with an experimental design, it predicts qMRI task performance. It does not give explicit insight into either (i) whether this performance is ‘good’ or (ii) how it might be improved. The pipeline’s use-cases highlighted throughout this thesis involve evaluating a range of user-supplied candidate experiments, from which the best is selected. *Optimising* these experiments, i.e. adjusting them to improve their performance, is beyond the pipeline’s scope. Chapter 5 shows how insight into manual optimisation strategies may be gained by moving beyond ‘summary’ task performance metrics, but this process is neither automated nor independent of the user’s choice of candidate experimental settings. These limitations could be addressed by incorporating the pipeline into a self-contained *optimisation* framework, in which experimental designs are automatically adjusted and improved.

The complex interactions between θ_{mod}, θ_{est}, and θ_{acq}, which themselves motivate the need for a holistic CED approach, render experimental optimisation challenging: the partial derivatives of task performance with respect to θ are analytically intractable. This sets the stage for a machine-learning-based optimisation approach, perhaps implemented within a reinforcement learning (RL) framework: the assessment pipeline presented in this thesis would act as the RL *environment*, the RL *agent* would adjust experimental design settings, leading to changes in RL *state* (task performance), used to guide subsequent algorithm iterations.

An ML optimization strategy would also motivate a significant adjustment in the CED pipeline’s structure. The simulation of parameter estimation and task evaluation could be *combined* and replaced by a new 'task-solving' network. This network would map input data (e.g. signal intensities) to task outputs (e.g. tissue classification), bypassing model selection, parameter estimation, and biomarker selection. It would be trained, in conjunction with the RL optimizer, to determine truly data-driven optimal experimental settings.

6.2.2 Pipeline scope

Throughout this Thesis, MR scanners have been treated as outputting spatially-mapped signal intensities, as per $S_{acquired}$ in Equation 2.3. Whilst appropriate for most 'off-the-shelf' qMRI experiments, this treatment conceals the time-series nature of acquired data (known as 'k-space'). Modern scanners perform extensive processing, controlled by user-selected hyperparameters, to convert this k-space data into outputted $S_{acquired}$. Developing the pipeline to simulate some, or all, of this in-scanner processing would enable the assessment and optimization of the associated hyperparameters and lead to further-improved qMRI CED.

6.2.3 Pipeline validation

Another avenue for future work is to expand the validation work described in Chapter 3. To demonstrate the reliability of in-silico task performance predictions, this thesis has relied on pre-existing in-vivo datasets not originally designed for CED validation studies. This has limited the scope of the validation to analysing differences in qMRI post-processing (θ_{mod} and θ_{est}) rather than data acquisition (θ_{acq}). Future validation studies could exploit bespoke in-vivo datasets, acquired with a wide range of θ_{acq}, to provide additional evidence that the assessment pipeline reliably predicts task performance across a wide range of candidate experimental designs.

6.2.4 Tunable parameter estimator

Along similar lines, there is significant scope for further testing and validation of the hybrid parameter estimator described in Chapter 4. This method defines a tunable training loss which enables a parameter estimation network to straddle the inter-estimator bias-variance trade-off discussed above. A proof-of-concept version of this method is presented, implemented for a single value of the tunable parameter α; following the general theme of this thesis, this work could be expanded by training a range of estimators, each with a different α, and repeat the tasks 'W1' and 'W2' described in Chapter 5 to find the estimator which maximises task performance.

6.2.5 Hyperparameter optimisation

The parameter estimation networks presented in Chapter 4 share many structural properties: number of neurons, number of layers, activation functions, optimization algorithms, et cetera. This commonality was a deliberate choice in the context of Chapter 4, which controlled as many inter-network differences as possible to isolate the effect of changing training loss. This shared structure was taken from the DNN qMRI literature; it is possible, indeed probable, that parameter estimation performance could be improved by network-specific structural optimization.

6.2.6 Patch-based parameter estimation

Chapter 4 reports bias-variance trade-offs in voxelwise DNN estimators; whether these trends extend to patch-based convolutional neural networks (CNNs) remains as yet unknown. CNNs map spatially-related clusters of qMRI signals to corresponding clusters of parameter estimates. Such networks are compatible with all DNN training approaches described in this thesis and present an obvious test-case for the generalisability of adjusting network training labels.

6.2.7 Computational tool utility

The highest-impact opportunity for further work lies in making the computational tools developed in this thesis *usable*, *accessible*, and *useful* to the people who actually design qMRI experiments. This work could follow two streams.

The first involves pure software engineering: adapting the codebase that underpins this thesis (publicly available, fully documented, here and here) into a software package usable by clinicians with limited technical expertise. The development of this package could incorporate an implementation of the RL-based optimization framework described above.

The second stream involves expanding the functionality of Chapter 2's assessment pipeline to make it more attractive to end-users. Possible extensions include: explicit calculation of acquisition time, allowing users to determine either minimum acquisition time for given task performance, or maximum performance for given acquisition time; an interactive 'simulator' which visually displays trade-offs between a range of pre-computed experimental setting combinations; and an extension from voxelwise simulation to whole-organ, whole-slice, or whole-volume imaging, providing a qualitative 'feel' for the kinds of visual contrast generated by different experimental designs.

6.2.8 Final thoughts

This thesis has made the case, and set the stage, for task-driven qMRI CED. Its central argument is simple: qMRI is a *tool* which should be *adjusted* to its intended use, whatever that may be. The contributions presented in Chapters 2-5 demonstrate various ways in which this adjustment may be achieved. This chapter has suggested avenues for further development and consolidation into useful, user-friendly tools. It is the author's hope that the work presented in this thesis will mark the first step towards qMRI experiments being routinely designed - and tailored - for *specific* experimental tasks.

Appendix A

Research output

A.1 Journal papers

1. **Epstein SC**, Bray TJP, Hall-Craggs MA, Zhang H, *Task-driven assessment of experimental designs in diffusion MRI: a computational framework*, 2021, PLOS ONE 16(10): e0258442

2. **Epstein SC**, Bray TJP, Hall-Craggs MA, Zhang H, *Choice of training label matters: how to best use deep learning for quantitative MRI parameter estimation*, 2022, arXiv:2205.05587, *under review*

3. Parker CS, Schroder A, **Epstein SC**, Cole J, Alexander DC, Zhang H, *Rician likelihood loss for quantitative MRI using self-supervised deep learning*, 2023, arXiv:2307.07072, *under review*

A.2 Conference abstracts

1. **Epstein SC**, Bray TJP, Hall-Craggs M and Zhang H, *Variability from complexity: assessing IVIM acquisition schemes through parameter estimation uncertainty*, 2020, ISMRM Annual Meeting 2020

2. **Epstein SC**, Bray TJP, Hall-Craggs M and Zhang H, *Towards a computational framework for task-driven experimental design*, 2021, ISMRM Annual Meeting 2021

3. **Epstein SC**, Bray TJP, Hall-Craggs M and Zhang H, *Quantitative MRI parameter estimation with supervised deep learning: MLE-derived labels outperform groundtruth labels*, 2022, ISMRM Annual Meeting 2022

4. **Epstein SC**, Bray TJP, Hall-Craggs M and Zhang H, *Do deep learning-based qMRI parameter estimators improve clinical task performance?*, 2023, ISMRM Annual Meeting 2023

5. Guerreri M, **Epstein S**, Azadbakht H, and Zhang H, *Can machine learning resolve model degeneracy in tissue microstructure estimation?*, 2023, ISMRM Annual Meeting 2023

A.3 Conference proceedings

1. Lim JP, Blumberg SB, Narayan N, **Epstein SC**, Alexander DC, Palombo M, Slator PJ, *Fitting a Directional Microstructure Model to Diffusion-Relaxation MRI Data with Self-supervised Machine Learning*, 2022, Computational Diffusion MRI - 13th International Workshop

2. Guerreri M, **Epstein SC**, Azadbakht H, Zhang H, *Resolving quantitative MRI model degeneracy with machine learning via training data distribution optimisation*, 2023, 28th Biennial Information Processing in Medical Imaging (IPMI 2023)

Appendix B

Cramér-Rao Lower Bound (CRLB)

The Cramér-Rao Lower Bound (CRLB) provides a surrogate measure for biomarker variance arising from stochastic noise.

Consider the set of all unbiased estimators which, under noise-free conditions, are able to perfectly invert a qMRI forward model to reconstruct ground-truth biomarker estimates. Once noise is introduced to the system, such estimators will necessarily produce some variance in their estimates. The CRLB provides the minimum noise-induced variance achievable by *any* of these estimators.

Specifically, the CRLB is derived from the Fisher Information matrix (FIM), which itself is a measure of the information that an observable variable (MRI signal) contains about a latent variable upon which it depends (qMRI biomarker). The FIM contains information about how the log-likelihood of an observed signal varies with the underlying latent variables; it is defined as the variance of the score S:

$$S = \frac{\partial L(X;\theta)}{\partial \theta} \tag{B.1}$$

where $L(X;\theta)$ is the log-likelihood of X (observed MRI signal) given θ (qMRI biomarkers) under noise. The FIM ($\mathcal{I}$) is the variance of the score, evaluated at the maximum likelihood estimate, i.e. at the true, unbiased estimate of θ:

$$\mathcal{I}_{i,j} = E[(S-\bar{S})^2|\theta] = E[S^2|\theta] = -E\left[\frac{\partial^2 L(X;\theta)}{\partial\theta_i \partial\theta_j}\middle|\theta\right] \tag{B.2}$$

where E represents the expectation and $E[S|\theta] = 0$ [94, p. 116].

The FIM gives the curvature of the log-likelihood surface at the maximum likelihood estimate. If $\mathcal{I}_{i,j}$ is large, the log-likelihood varies rapidly with θ_i, θ_j, i.e. $L(X; \theta \neq \theta_{MLE}) \sim L(X; \theta_{MLE})$ for a small range of θ: there is low uncertainty in estimates of the latent variable θ. Conversely, if $\mathcal{I}_{i,j}$ is low, there is a large range θ_i, θ_j for which $L(X; \theta \neq \theta_{MLE}) \sim L(X; \theta_{MLE})$, and there is high uncertainty in estimates of θ.

The CRLB is defined as the inverse of the FIM:

$$Var(\hat{\theta}_{i,j}) \geq \frac{1}{\mathcal{I}_{i,j}} \equiv CRLB \tag{B.3}$$

A large $\mathcal{I}_{i,j}$ encodes a large log-likelihood curvature at the maximum likelihood estimate and, consequently, a low minimum variance achievable by a perfect unbiased estimator.

Appendix C

Diffusion MRI and the Intravoxel Incoherent Motion model

Contents

Diffusion MRI (dMRI) is a form of quantitative magnetic resonance imaging (qMRI) which exploits the self-diffusive properties of water molecules. Self-diffusion describes the random-walk Brownian motion of molecules within a fluid in the absence of (i) a chemical potential gradient and (ii) external interactions.

dMRI relies on the fact that as water molecules self-diffuse through *tissue*, they are in fact *not* free of external interactions; their motion is *restricted* by intracellular structures ("microstructure"), and their self-diffusion is limited. This self-diffusive disruption encodes information about tissue microstructure, and dMRI experiments are designed to capture this information.

dMRI cannot, unfortunately, track the random-walk motion of *individual* water molecules. Rather, the dMRI signal S is a *sum* of the signals S_i of all the N molecules within a voxel[1]:

$$S = \sum_{i}^{N} S_i \tag{C.1}$$

dMRI experiments encode microstructural information in S by introducing molecular-motion-dependant *phase terms* to the constituent S_i. As individual molecular paths

[1] For simplicity, 'water molecule' is used interchangeably with H proton; it is assumed that any signal arising from fat has been suppressed, as is common in dMRI acquisition.

diverge over time, so does their accrued phase; the resulting intra-molecular incoherence attenuates the measured signal S.

Consider the signal arising from a voxel containing N water molecules following random-walk trajectories $r_i(t)$. Suppose we apply two *sequential* spatially-dependant phase terms to each molecule's signal S_i:

$$S = \sum_{i}^{N} S_i \int_{0}^{\alpha} e^{i\phi(r_i(t))} dt \int_{\alpha+\beta}^{2\alpha+\beta} e^{i-\phi(r_i(t))} dt \tag{C.2}$$

where α and β are user-selected time constants, such that the two phase accruals (a) have the same spatial dependence and (b) are equal in magnitude but opposite in sign.

Stationary molecules (constant $r_i(t)$) are unaffected by this sequence: the first phase addition (ϕ) is cancelled out by the subsequent phase removal ($-\phi$). In contrast, the S_i of self-diffusing molecules acquires a *net phase* which encodes the random path $r(0 \leqslant t \geqslant 2\alpha + \beta)$ they have taken.

The summation that underpins the voxelwise signal S is *attenuated* by the magnitude and relative incoherence of the phase acquired by each constituent water molecule. In the simplest terms, the more 'restrictive' the microstructure within a voxel, the shorter the random-walk of the associated water molecules, the smaller the (and more coherent) the acquired phase, and the smaller the attenuation.

C.1 Apparent Diffusion Coefficient

This relationship forms the basis of the apparent diffusion coefficient (ADC) dMRI model:

$$\frac{S(b)}{S_0} = e^{-bD_{ADC}} \tag{C.3}$$

where $b(\alpha, \beta, \phi)$, the *diffusion weighting*, is an independent variable which relates signal attenuation to the molecular trajectories $r(t)$; D_{ADC} is a measure of water *diffusivity* which depends on the magnitude and coherence of the phase accrued across a voxel of interest; and S_0 is the signal measured for $b = 0$, i.e. zero phase accrual. An ADC dMRI experiment consists of acquiring $S(b)$ at multiple b-values, and fitting the model parameters (S_0 and D_{ADC}) to the resulting data. Larger best-fit ADC estimates ($\overline{D_{ADC}}$) correspond to faster signal attenuation and imply less restricted diffusion.

C.2 Intravoxel Incoherent Motion

The microstructure that dMRI experiments probe is micrometer-scale, whilst dMRI voxels are millimeters across. Each voxel may therefore contain a wide range of *different* microstructural environments, each restricting water molecules in its own unique way. The ADC model averages out these differences by assuming that all molecules in voxel can be described by the *same* D_{ADC}.

More complex dMRI models address this simplification by introducing the concept of *signal compartments*. The single voxelwise summation described in Equation C.1 is decomposed into multiple summations:

$$S = \sum_{j}^{n} \sum_{i}^{N} S_{i,j} \tag{C.4}$$

where j refers to a signal compartment which groups water molecules experiencing similar microstructural environments. A simple extension of the ADC model into a multi-compartment one results in:

$$\frac{S(b)}{S_0} = \sum_{j}^{n} \gamma_j e^{-bD_j} \tag{C.5}$$

where D_j is a measure of the diffusivity of the molecules in the jth compartment, and the volume fraction γ_j describes the proportion of the voxel's water molecules that experience the jth compartment's microstructural environment.

The intravoxel incoherent motion (IVIM) model [75], referred to throughout this thesis, is a two-compartment ($n = 2$) extension of ADC which accounts for a *non-diffusive* source of motion detected by dMRI experiments:

$$\frac{S(b)}{S_0} = \gamma_1 e^{-bD_1} + \gamma_2 e^{-bD_2} \tag{C.6}$$

The first compartment ($j = 1$) is equivalent to ADC: voxelwise averaging of all self-diffusive incoherence effects, summarised by diffusivity D_1.

The second compartment ($j = 2$) models microcirculation: water which not only self-diffuses, but also *flows* in the capillary network within the voxel of interest. The capillaries that form this network are oriented quasi-randomly; as water molecules move along them, their path r_t resembles Brownian motion, albeit much faster than what arises from self-diffusion. Much like in ADC, this motion leads to exponential attenuation described by $D_2 \gg D_1$.

To reflect this biophysical insight, throughout this thesis Equation C.6 has been rewritten as:

$$\frac{S(b)}{S_0} = fe^{-b(D_{fast}+D_{slow})} + (1-f)e^{-bD_{slow}} \tag{C.7}$$

where the perfusion fraction f (i.e. γ_2) refers to the fraction of water molecules in circulating in the capillary network; D_{slow} (i.e. D_1) refers to the mean diffusivity of non-microcirculating water molecules; and D_{fast} (i.e. $D_2 - D_1$) refers to the pseudo-diffusivity attributed to capillary-mediated translation motion.

Appendix D

A brief introduction to deep neural networks used in qMRI parameter estimation

Contents

The deep learning (DL) parameter estimators discussed throughout this thesis are DNNs. This appendix provides a brief introduction to qMRI parameter estimation DNNs, intended to provide some basic theoretical background to Chapter 4.

D.1 Anatomy of a DNN

Deep neural networks $\mathcal{N}$ are, in general, many-to-one functions:

$$\mathcal{N}(x|\phi, w) = y \tag{D.1}$$

which, parameterised by ϕ and w, map inputs x to outputs y.

These functions are generated by combining small building block functions ($\psi(x|w)$), known as *neurons*. $\mathcal{N}$'s *structure* parameter ϕ describes the form and arrangement of these neurons, whilst the *weighting* parameter w describes the constituent neurons' own parameterisation.

Consider a simple neuron $\psi(x|w)$:

$$\psi(x|w) = x + w \tag{D.2}$$

which returns its input x biased by a constant term w. Combining three such neurons in series produces a simple network:

$$\mathcal{N}(x|\phi, w) = \psi^{(3)}(\psi^{(2)}(\psi^{(1)}(x|w^{(1)})|w^{(2)})|w^{(3)}) = ((x + w^{(1)}) + w^{(2)}) + w^{(3)} \tag{D.3}$$

which adds 3 constants ($w^{(1)}$, $w^{(2)}$, $w^{(3)}$) to an input x. The structure parameter ϕ describes the number of neurons (three), their form (constant bias), and how they are combined (in series). The weighting parameter w describes the the constants being added by each neuron. This simple network can be visualised as a series of connected *layers*:

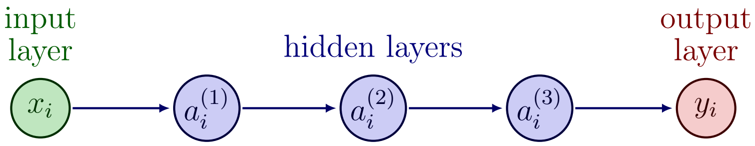

Figure D-1: A simple network.

where i refers to the i^{th} element of x and $a_i^{(1)} = \psi^{(1)}(x_i|w^{(1)}) = x_i + w^{(1)}$ is known as the *activity* of $\psi^{(1)}$ [95]. The network's input x is known as the *input layer*, the output y as the *output layer*, and the structure that connects them as the network's *hidden layers*. For $x, y \in \mathbb{R}^3$:

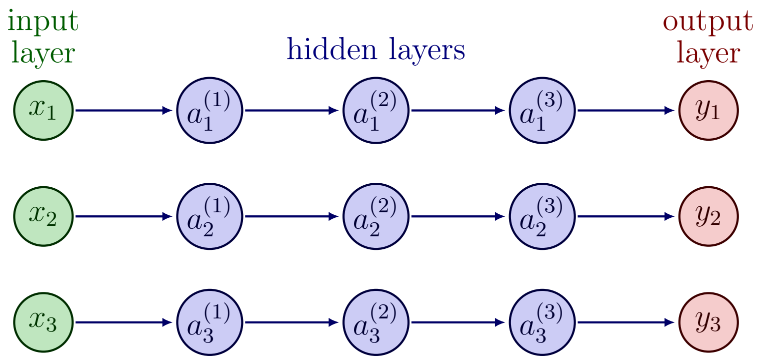

Figure D-2: A simple network of 'width' 3.

the network $\mathcal{N}$ adds $w^{(1)} + w^{(2)} + w^{(3)}$ to *each element* of x.

$\mathcal{N}$ encodes a *mapping*, from input-space X to output-space Y, which can be adjusted by changing the values of w. In this example, N can only encode addition, no matter the choice of w. The chosen structure parameter ϕ has severely limited this N's ability to encode more complex - and potentially useful - functions.

The parameter estimation DNNs described in Chapter 4, which encode complex non-linear functions, differ in two important ways from the toy example given above: (i) the nature of the neurons that underpin them and (ii) how these neurons are connected to each other.

D.2 Neurons

Neurons $\psi(x|w)$ convert inputs to activities; in the example above this consists of a simple bias operation applied to a single input. In contrast, the parameter estimation networks discussed in this thesis contain neurons which convert multiple inputs into single outputs. This is achieved by compositing three conceptually-distinct operations: weighted summation, bias, and rescaling.

D.2.1 Weighted summation

The first operation performed by each neuron is computing a weighted sum of its inputs:

$$\alpha^j_{intermediate} = \sum_{i=1}^{I} \omega^j_i a^{j-1}_i \tag{D.4}$$

where a^{j-1} is the j^{th} neuron's input (i.e. the *output* of the j-1$^{\text{th}}$ neuron) and I is its dimensionality, indexed by i. The weights ω_i^j are adjustable parameters which, much like the constants $w^{(1)}$-$w^{(3)}$ above, control neuron's behaviour.

D.2.2 Bias

The second operation adds a bias b to the weighted sum computed above:

$$\alpha^j = \alpha^j_{intermediate} + b^j \tag{D.5}$$

and yields α^j, the j^{th} neuron's *activation*. The bias b is, just like w_i^j, an adjustable parameter which controls each neuron's performance.

D.2.3 Rescaling

The final operation involves rescaling the activation α^j to generate the *activity* a^j:

$$a^j = f(\alpha^j) \tag{D.6}$$

where f, known as the *activation function*, is user-defined and, unlike w, generally not adjusted during training (see below). Activation functions introduce non-linearity into $\psi(x|w)$, enabling N to represent *non-linear functions* of x.

DNNs employ a wide range of activation functions; the parameter estimators implemented in Chapter 4 employ the exponential linear unit activation (ELU) activation function [89]:

$$\text{ELU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \gamma(e^x - 1), & \text{otherwise} \end{cases} \tag{D.7}$$

D.2.4 Bringing it all together

Combining these three operations gives an overall definition of the neuron $\psi(x|w)$:

$$\psi(x|w) = a(x|w) = f(b + \sum_{i=1}^{I} \omega_i x_i | \gamma) \tag{D.8}$$

where the network weights w describe β, γ, and ω[1].

D.3 Connections

DNNs are composed of large numbers of such neurons. Individual neurons' inputs and outputs are linked by a *layered* network structure. Neurons occupying different layers may be *connected*, whereby the output of one neuron is provided as input to another. Neurons within the same layer are never connected to each other. The number of neurons within a layer is known as that layer's *width*. The number of layers within a network is known as that network's *depth*. Any layers lying between input and output are known as *hidden*; the term 'deep' in Deep Neural Network refers to DNNs having *depth* of more than one hidden layer.

The parameter-estimation DNNs described throughout this thesis are connected *sequentially* (layers are arranged in series, such that neuronal connections are only formed between adjacent layers) and *fully* (all neurons within a layer are connected to all other neurons in adjacent layers).

An example of such a network, similar to those described in Chapter 4, is visualised in Figure D-3.

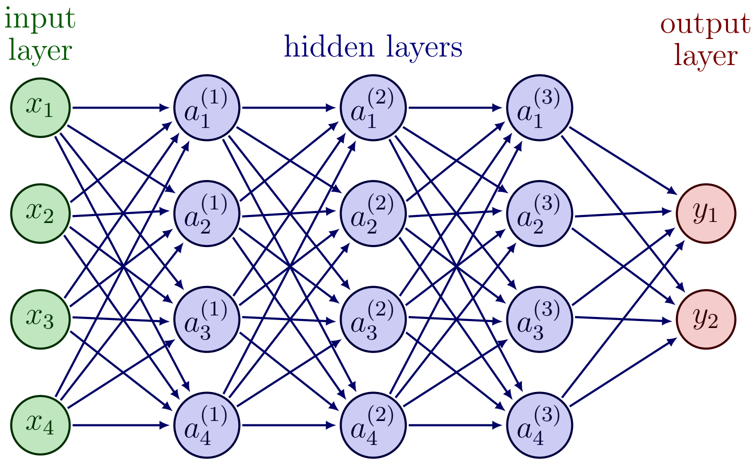

Figure D-3: A basic parameter estimator.

This network takes a four-dimensional input (x); each dimension (x_1 to x_4) is passed to each of the four neurons ($a_1^{(1)}$ to $a_4^{(1)}$) in the first hidden layer. These neurons are each parameterised by 5 tunable parameters: four weighting terms ω and one bias term b.

[1] Note that, for simplicity, α is here subsumed by the weighting parameter w rather than, as is more common, the structure parameter ϕ.

The first hidden layer generates four outputs (activities a_1^1 to a_4^1) which are passed to each of the four neurons in the *second* hidden layer. This process is repeated for the third, and final, hidden layer. The output layer contains two neurons. Each of these neurons computes its activation in the same way as the hidden layers that precede it: weighted summation, biasing, rescaling.

In this way, this network combines 14 non-linear functions (three hidden layers of width 4, plus one output layer of width 2), parameterised by 70 tunable terms[2] (14 neurons, each controlled by 4 weight terms and 1 bias), to convert a four-dimensional input into a two-dimensional output.

This network 'becomes' a qMRI parameter estimator if (i) it is provided qMRI signals as inputs x and (ii) its outputs y are *interpreted* as qMRI model parameters. Using the notation from Section 2.1, the network $\mathcal{N}$ becomes:

$$\mathcal{N}(S_{acquired}) = \overline{P} \tag{D.9}$$

The quality of this estimator - that is, how well it maps signals to meaningful parameter estimates - depends on both its structure ϕ (i.e. its *capacity* to encode this mapping) and its weights w (i.e. its tuning to this *specific* mapping).

The remainder of this appendix considers cases where network performance is not limited by structure parameter ϕ, but rather depends exclusively on the tunable network weights w. Network training attempts to determine the optimal w.

D.4 Training & inference

DNN training is the process by which $\mathcal{N}$'s weighting parameter w is adjusted to maximise network *performance*. In the context of qMRI, using notation introduced in Chapter 2, a maximally-performing network is one which converts qMRI signals $S_{acquired}$ into desired parameter predictions P_{target}.

More generally, such a network can be expressed as:

$$\mathcal{N}(x|\phi, w_{target}) = y_{target} \tag{D.10}$$

where w_{target} are the network weights which generate 'ideal' predictions y_{target} from network inputs x.

[2] Note that the scaling factor γ is not included in this count. It is normally considered a structure parameter ϕ as it is kept constant during training.

The training process approximates w_{target} with *learned* weights w_{train} by exploiting a *training set* of network inputs (x_{train}). Each member of this training set is associated with an 'ideal' network output y_{target}, either implicitly (e.g. *Self-supervised* in Chapter 4) or explicitly (e.g. $Supervised_{GT}$, also in Chapter 4).

Each element of x_{train} is input to the network, generating outputs y_{train}. A loss function L, which evaluates the distance between y_{train} and y_{target}, is then iteratively minimised over the entire training set:

$$w_{train} = \arg\min_{w} \sum_{x_{train}} L(y_{train}, y_{target}|w) \tag{D.11}$$

The weights w_{train} obtained by this process parameterise the network which, when evaluated over the entire training set, generates predictions which are closest to the supplied training data's 'ideal' network outputs.

The network is now trained, and ready for *inference*: application to unseen data. With an appropriate choice of x_{train}, the learned parameters w_{train} encode a mapping which *generalises* and produces useful outputs for inputs not encountered during training.

The mechanics of the optimisation procedure described in Equation D.11 are beyond the scope of this Appendix. For the purposes of Chapter 4, what *is* important is that network weights - and therefore network predictions - depend heavily on both how L is *formulated*, and on how y_{target} is defined.

Appendix E

In-vivo parameter estimation validation strategies

Contents

The process of assessing parameter estimators can be thought of as involving three steps: (i) posing an ‘assessment question’, (ii) defining a metric that answers this question, and finally (iii) describing a way to evaluate this metric.

When using in-silico qMRI test data, which has been simulated from known generative groundtruths, an example of this three-step process might be:

1. Posing an ‘assessment question’

    - Does the parameter estimator provide *accurate* estimates of the generative qMRI parameters?

2. Defining a metric that answers this question
    - Mean difference between groundtruth and estimate.
3. Describing a way to evaluate this metric
    - Calculate parameter estimates for entire dataset; calculate pairwise difference between groundtruth and estimate; compute mean of these differences.

Unlike synthetic data, in-vivo datasets do not contain ‘groundtruth’ values usable as reference during assessment. What follows is a list of existing in-vivo qMRI assessment strategies which attempt to bypass this limitation, compiled here for reference.

E.1 Additional data (different modality)

These methods supplement their qMRI test data with non-MRI datasets (associated with the same tissue(s)) which provide independent measurements of qMRI biomarkers.

E.1.1 Reference for comparison: histology (ex-vivo)

Assessment question: Do qMRI parameter estimates match independent measures of tissue properties?

Rationale: Agreement suggests information reliability.

Assessment criterion: Qualitative agreement [79], Spearman’s rank [96].

Example implementation: Acquire qMRI data → dissect tissue → measure qMRI biomarker from histology → register histology to qMRI data

E.2 Additional data (same modality)

These methods supplement their qMRI test data with additional qMRI data acquisitions of the same tissue(s).

E.2.1 Reference for comparison: super-sampled dataset

Assessment question: Do real-world parameter estimates match those obtained when acquisition setting limitations are removed/reduced?

Rationale: Super-sampled acquisitions are information-rich and, combined with maximum likelihood estimation, provide reliable reference for comparison.

Assessment criterion: RMSE between estimates and super-sample-derived MLE estimates [58, 97, 98, 99].

Example implementation: Acquire supersampled data → calculate MLE parameter estimates → sub-sample data to match θ_{acq} of interest → compare sub-sampled estimates with super-sampled reference

E.2.2 Self consistency: super-sampled dataset

Assessment question: Are parameter estimates robust to noise and sampling scheme?

Rationale: Reliable techniques should give consistent parameter estimates.

Assessment criterion: Variance [100], ROI mean agreement [101], estimated signal agreement via Bland-Altmann plots and correlation coefficients [101].

Example implementation: Acquire supersampled data → randomly subsample data to match θ_{acq} of interest → compare parameter estimates across subsamples

E.2.3 Self consistency: repeat acquisition

Assessment question: Are parameter estimates robust to noise?

Rationale: Reliable techniques should give consistent parameter estimates.

Assessment criterion: Median voxelwise standard deviation [98], ROI mean agreement [102]

Example implementation: Acquire multiple acquisitions of the same tissue → compare parameter estimates across acquisitions

E.3 No additional data

These methods forego independent references for comparison, and therefore do not require supplementary test data.

E.3.1 Self consistency: intra-subject, intra-ROI

Assessment question: Are parameter estimates consistent across similar tissues?

Rationale: Reliable techniques should give consistent parameter estimates.

Assessment criterion: Ratio of mean parameter value to standard deviation [85, 96, 101, 103]

Example implementation: Draw ROIs of homogeneous anatomies → compare parameter estimate consistency within each ROI.

E.3.2 Task performance: intra-subject, inter-ROI

Assessment question: Are parameter estimates sensitive to tissue differences?

Rationale: Useful techniques must be sensitive to changes in underlying tissue properties.

Assessment criterion: Percentage difference [102, 103], Student's t-test [60], ROC curves [98].

Example implementation: Draw ROIs of different anatomies within subjects → compare parameter estimates between different ROIs.

E.3.3 Self consistency: inter-subject, inter-ROI

Assessment question: Are similar-tissue parameter estimates consistent across subjects/acquisitions?

Rationale: Reliable techniques should give consistent parameter estimates.

Assessment criterion: intraclass correlation coefficient of ROI mean parameter values for a range of matched tissues across multiple subjects [61].

Example implementation: Draw ROIs of equivalent anatomies across multiple subjects → compare matched parameter estimates.

E.3.4 Intrinsic quality assessment

Assessment question: Do parameter estimates well-represent the acquired signal?

Rationale: Parameter estimates should describe the signal they are obtained from.

Assessment criterion: RMSE signal residuals [60], correlation between model parameters (lower is better, such that each parameter encodes unique information) [102], visual assessment of parameter map [100].

Example implementation: Evaluate parameter estimates → assess stand-alone quality of fit.

E.3.5 Similarity assessment

Assessment question: Do parameter estimates agree with those generated by another method?

Rationale: Cheap methods can replace more expensive ones if they generate similar parameter estimates.

Assessment criterion: Student's t-test [60].

Example implementation: Evaluate parameter estimates using multiple methods → compare results across methods.

E.3.6 Network training robustness

Assessment question: Are machine learning methods robust to changes in training and testing data?

Rationale: Reliable techniques should not depend stochastically on selection of training/testing data.

Assessment criterion: Signal residuals [98].

Example implementation: Train networks repeatedly, each time randomising the split between training and testing data → compare quality of fit across repetitions.

Appendix F

Notation describing qMRI acquisition

Section 2.1 introduces, in general terms, notation used to describe qMRI experiments; this Appendix grounds this notation by providing a concrete example: acquiring dMRI parameter maps of subchondral bone marrow. This exemplar experiment consists of (i) acquiring 10 spatial maps ('images') of the tissue, each associated with a distinct diffusion weighting ('b-value' [4, p. 624]) and (ii) fitting the IVIM signal model to each spatial location shared across these images, generating three dMRI parameter maps (one for each IVIM parameter).

What follows is a reproduction of the key equations found in Section 2.1; in each case, new notation is related to the simple experiment described above.

$$S_{deterministic} = \mathcal{M}(\theta_{acq}, T) \approx M(\theta_{acq}, P) \tag{F.1}$$

- $S_{deterministic}$ is a 10-by-1 vector containing voxel intensities across the 10 recorded b-values. This vector is deterministic - and is never measured directly - as it precedes the addition of noise inherent to the imaging system.

- $\mathcal{M}$ is a function which encodes the deterministic relationship between acquisition settings θ_{acq} and tissue properties T. This function underpins the generation of voxel intensities and its form is, in general, unknown.

- θ_{acq} is the set of experimental instructions which encode the acquisition of qMRI data; in this example, this subsumes the 10 diffusion-weighted b-values, the properties of the scanning hardware (make, model, coils used, etc.) as well as software (acquisition program, gradient timings, directions, etc.). θ_{acq} is the known independent variable in data acquisition.

- T is the set of tissue properties which, via interactions amongst themselves and with the scanner, lead to differences in voxel intensities; in this example, this subsumes the number and distribution of protons within the sample, its cellular structure, chemical environment(s), etc.; T is, in the case of biological tissue, unknown.

- M is a function, known as the signal model, which encodes the deterministic relationship between θ_{acq} and its parameters P. It predicts voxel intensities and, in this example, is the IVIM model.

- P is a 3-by-1 vector which describes the values taken by the 3 parameters of the IVIM model.

$$S_{acquired} = \mathcal{M}(\theta_{acq}, T) + \epsilon \equiv S_{deterministic} + \epsilon \tag{F.2}$$

- ϵ is a 10-by-1 vector containing noise instantiations, drawn from a Rician distribution, associated with the 10 voxel intensities $S_{deterministic}$.

$$P_{target} = \Phi(T; \theta_{est}, \theta_{mod}, \theta_{acq}) \tag{F.3}$$

- θ_{est} is the set of experimental instructions which encode the process of parameter estimation; in this example, as in Chapter 2, this may describe bcNLLS and the hyperparameters that determine its implementation.

- θ_{mod} is the set of experimental instructions which encode the choice of signal model; this may in general vary across an image (on an anatomy-by-anatomy basis) but in this example describes the IVIM model for all voxels.

- P_{target} is a 3-by-1 vector which describes an experimental designer's 'ideal' parameter estimates for a given tissue T and experimental settings θ; in this example, this may correspond to IVIM parameters which give low-bias approximations of the bone marrow's perfusion fraction[1].

- Φ is a function which maps tissue properties T to corresponding user-defined 'ideal' parameter estimates P_{target}, *for a given experimental design*. When dealing with in-vivo data, this function has poorly-defined inputs (T is unknown) and cannot be computed directly. In such cases it is approximated by the methods described in Appendix E.

$$\arg\min_{\overline{P}} \mathcal{D}(\overline{P}; P_{target}) \tag{F.4}$$

[1] See Appendix C for a description of this signal model M.

- $\overline{P}$ is a 3-by-1 vector which describes a specific P: the best-fit parameter estimates taken by the 3 parameters of the IVIM model.

Bibliography

[1] Concepción González Hernando, Laura Esteban, Teresa Cañas, Enrique Van den Brule, and Miguel Pastrana. The role of magnetic resonance imaging in oncology. *Clinical and Translational Oncology*, 12(9):606–613, 2010.

[2] Maythem Saeed, Tu Anh Van, Roland Krug, Steven W. Hetts, and Mark W. Wilson. Cardiac mr imaging: current status and future direction. *Cardiovascular Diagnosis and Therapy*, 5(4), 2015.

[3] M Symms. A review of structural magnetic resonance neuroimaging. *Journal of Neurology, Neurosurgery &amp Psychiatry*, 75(9):1235–1244, 2004.

[4] E.M. Haacke, R.W. Brown, M.R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley, 1999.

[5] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69(3):89–95, 2001.

[6] George H. Weiss, Raj K. Guptaj, James A. Ferretti, and Edwin D. Becker. The choice of optimal parameters for measurement of spin-lattice relaxation times. I. Mathematical formulation. *Journal of Magnetic Resonance (1969)*, 37(3):369–379, feb 1980.

[7] Henry Z. Wang, Stephen J. Riederer, and James N. Lee. Optimizing the precision in T1 relaxation estimation using limited flip angles. *Magnetic Resonance in Medicine*, 5(5):399–416, 1987.

[8] Alex D. Bain. The choice of parameters in an NMR experiment. Application to the inversion-recovery T1 method. *Journal of Magnetic Resonance (1969)*, 89(1):153–160, 1990.

[9] J. A. Jones, P. Hodgkinson, A. L. Barker, and P. J. Hore. Optimal sampling strategies for the measurement of spin-spin relaxation times. *Journal of Magnetic Resonance - Series B*, 113(1):25–34, 1996.

[10] Yantian Zhang, Hong N. Yeung, Matthew O'Donnell, and Paul L. Carson. Determination of sample time for T1 measurement. *Journal of Magnetic Resonance Imaging*, 8(3):675–681, 1998.

[11] Mehmet Akçakaya, Sebastian Weingärtner, Sébastien Roujol, and Reza Nezafat. On the selection of sampling points for myocardial T1 mapping. *Magnetic Resonance in Medicine*, 73(5):1741–1753, 2015.

[12] Gopal Nataraj, Jon Fredrik Nielsen, and Jeffrey A. Fessler. Optimizing MR scan design for model-based T1,T2 estimation from steady-state sequences. *IEEE Transactions on Medical Imaging*, 36(2):467–477, 2017.

[13] Rui Pedro A.G. Teixeira, Shaihan J. Malik, and Joseph V. Hajnal. Joint system relaxometry (JSR) and Crámer-Rao lower bound optimization of sequence parameters: A framework for enhanced precision of DESPOT T1 and T2 estimation. *Magnetic Resonance in Medicine*, 79(1):234–245, 2018.

[14] Anastasios Anastasiou and L. D. Hall. Optimisation of T2 and M0 measurements of bi-exponential systems. *Magnetic Resonance Imaging*, 22(1):67–80, 2004.

[15] Adrienne N. Dula, Daniel F. Gochberg, and Mark D. Does. Optimal echo spacing for multi-echo imaging measurements of Bi-exponential T2 relaxation. *Journal of Magnetic Resonance*, 196(2):149–156, 2009.

[16] Mara Cercignani and Daniel C. Alexander. Optimal acquisition schemes for in vivo quantitative magnetization transfer MRI. *Magnetic Resonance in Medicine*, 56(4):803–810, 2006.

[17] Marco Battiston, Francesco Grussu, Andrada Ianus, Torben Schneider, Ferran Prados, James Fairney, Sebastien Ourselin, Daniel C. Alexander, Mara Cercignani, Claudia A.M. Gandini Wheeler-Kingshott, and Rebecca S. Samson. An optimized framework for quantitative magnetization transfer imaging of the cervical spinal cord in vivo. *Magnetic Resonance in Medicine*, 79(5):2576–2588, 2018.

[18] Oscar Brihuega-Moreno, Frank P. Heese, and Laurance D. Hall. Optimization of diffusion measurements using Cramer-Rao lower bound theory and its application to articular cartilage. *Magnetic Resonance in Medicine*, 50(5):1069–1076, 2003.

[19] Daniel C. Alexander. A general framework for experiment design in diffusion MRI and its application in measuring direct tissue-microstructure features. *Magnetic Resonance in Medicine*, 60(2):439–448, 2008.

[20] D.H.J. Poot, A.J. den Dekker, E. Achten, M. Verhoye, and J. Sijbers. Optimal Experimental Design for Diffusion Kurtosis Imaging. *IEEE Transactions on Medical Imaging*, 29(3):819–829, mar 2010.

[21] Nienke D. Sijtsema, Steven F. Petit, Dirk H. J. Poot, Gerda M. Verduijn, Aad Lugt, Mischa S. Hoogeman, and Juan A. Hernandez-Tamames. An optimal acquisition and post-processing pipeline for hybrid IVIM-DKI in head and neck. *Magnetic Resonance in Medicine*, 85(2):777–789, 2021.

[22] Óscar Peña-Nogales, Diego Hernando, Santiago Aja-Fernández, and Rodrigo de Luis-Garcia. Determination of optimized set of b-values for Apparent Diffusion Coefficient mapping in liver Diffusion-Weighted MRI. *Journal of Magnetic Resonance*, 310:106634, jan 2020.

[23] Zohir Laib, Farid Ahmed Sid, Karim Abed-Meraim, and Aziz Ouldali. Estimation error bound for GRAPPA diffusion-weighted MRI. *Magnetic Resonance Imaging*, 74:181–194, 2020.

[24] Mojtaba Mirdrikvand, Harm Ridder, Jorg Thöming, and Wolfgang Dreher. Diffusion weighted magnetic resonance imaging for temperature measurements in catalyst supports with an axial gas flow. *Reaction Chemistry and Engineering*, 4(10):1844–1853, 2019.

[25] Björn Lampinen, Filip Szczepankiewicz, Danielle van Westen, Elisabet Englund, Pia C Sundgren, Jimmy Lätt, Freddy Ståhlberg, and Markus Nilsson. Optimal experimental design for filter exchange imaging: Apparent exchange rate measurements in the healthy brain and in intracranial tumors. *Magnetic resonance in medicine*, 77(3):1104–1114, 2017.

[26] Mario Sansone, Roberta Fusco, and Antonella Petrillo. D-optimal design of b-values for precise intra-voxel incoherent motion imaging. *Biomedical Physics and Engineering Express*, 5(3):035025, 2019.

[27] Sean P. Devan, Xiaoyu Jiang, Francesca Bagnato, and Junzhong Xu. Optimization and numerical evaluation of multi-compartment diffusion MRI using the spherical mean technique for practical multiple sclerosis imaging. *Magnetic Resonance Imaging*, 74:56–63, 2020.

[28] Paddy J. Slator, Jana Hutter, Andrada Ianus, Eleftheria Panagiotaki, Mary A. Rutherford, Joseph V. Hajnal, and Daniel C. Alexander. A Framework for Calculating Time-Efficient Diffusion MRI Protocols for Anisotropic IVIM and An Application in the Placenta. In *Computational Diffusion MRI. MICCAI 2019*, pages 251–263. Springer, Cham, 2019.

[29] Ivana Drobnjak and Daniel C. Alexander. Optimising time-varying gradient orientation for microstructure sensitivity in diffusion-weighted MR. *Journal of Magnetic Resonance*, 212(2):344–354, 2011.

[30] Benjamin Leporq, Hervé Saint-Jalmes, Cecile Rabrait, Frank Pilleul, Olivier Guillaud, Jérôme Dumortier, Jean-Yves Scoazec, and Olivier Beuf. Optimization of intra-voxel incoherent motion imaging at 3.0 Tesla for fast liver examination. *Journal of magnetic resonance imaging : JMRI*, 41(5):1209–17, 2015.

[31] Oscar Jalnefjord, Mikael Montelius, Göran Starck, and Maria Ljungberg. Optimization of b-value schemes for estimation of the diffusion coefficient and the perfusion fraction with segmented intravoxel incoherent motion model fitting. *Magnetic resonance in medicine*, 82(4):1541–1552, 2019.

[32] J. Choi and L.G. Raguin. Robust optimal design of diffusion-weighted magnetic resonance experiments for skin microcirculation. *Journal of Magnetic Resonance*, 206(2):246–254, 2010.

[33] Shantanu Majumdar, David C. Zhu, Satish S. Udpa, and L. Guy Raguin. A diffusion gradient optimization framework for spinal cord diffusion tensor imaging. *Magnetic Resonance Imaging*, 29(6):789–804, 2011.

[34] Philip K. Lee, Lauren E. Watkins, Timothy I. Anderson, Guido Buonincontri, and Brian A. Hargreaves. Flexible and efficient optimization of quantitative sequences using automatic differentiation of Bloch simulations. *Magnetic Resonance in Medicine*, 82(4):1438–1451, 2019.

[35] Vincent Gras, Ezequiel Farrher, Farida Grinberg, and N. Jon Shah. Diffusion-weighted DESS protocol optimization for simultaneous mapping of the mean diffusivity, proton density and relaxation times at 3 Tesla. *Magnetic Resonance in Medicine*, 78(1):130–141, 2017.

[36] Mustapha Bouhrara and Richard G. Spencer. Fisher information and Cramér-Rao lower bound for experimental design in parallel imaging. *Magnetic Resonance in Medicine*, 79(6):3249–3255, 2018.

[37] Paddy J. Slator, Jana Hutter, Laura McCabe, Ana Dos Santos Gomes, Anthony N. Price, Eleftheria Panagiotaki, Mary A. Rutherford, Joseph V. Hajnal, and Daniel C. Alexander. Placenta microstructure and microcirculation imaging with diffusion MRI. *Magnetic Resonance in Medicine*, 80(2):756–766, aug 2018.

[38] Da Xing, Nikolaos G. Papadakis, Christopher L.H. Huang, Vee Meng Lee, T. Adrian Carpenter, and Laurance D. Hall. Optimised diffusion-weighting for measurement of apparent diffusion coefficient (ADC) in human brain. *Magnetic Resonance Imaging*, 15(7):771–784, 1997.

[39] P.A. Armitage and M.E. Bastin. Utilizing the diffusion-to-noise ratio to optimize magnetic resonance diffusion tensor acquisition strategies for improving measurements of diffusion anisotropy. *Magnetic Resonance in Medicine*, 45(6):1056–1065, 2001.

[40] Khader M. Hasan, Dennis L. Parker, and Andrew L. Alexander. Comparison of gradient encoding schemes for diffusion-tensor MRI. *Journal of Magnetic Resonance Imaging*, 13(5):769–780, 2001.

[41] D.K. Jones, M.A. Horsfield, and A. Simmons. Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. *Magnetic Resonance in Medicine*, 42(3):515–525, 1999.

[42] Ivan Jambor, Harri Merisaari, Hannu J. Aronen, Jukka Järvinen, Jani Saunavaara, Tommi Kauko, Ronald Borra, and Marko Pesola. Optimization of b-value distribution for biexponential diffusion-weighted MR imaging of normal prostate. *Journal of Magnetic Resonance Imaging*, 39(5):1213–1222, 2014.

[43] Harri Merisaari and Ivan Jambor. Optimization of b-value distribution for four mathematical models of prostate cancer diffusion-weighted imaging using b values up to 2000 s/mm2: Simulation and repeatability study. *Magnetic Resonance in Medicine*, 73(5):1954–1969, 2015.

[44] Andreas Lemke, Bram Stieltjes, Lothar R. Schad, and Frederik B. Laun. Toward an optimal distribution of b values for intravoxel incoherent motion imaging. *Magnetic Resonance Imaging*, 29(6):766–776, 2011.

[45] Kishor Karki, Geoffrey D Hugo, John C Ford, Kathryn M Olsen, Siddharth Saraiya, Robert Groves, and Elisabeth Weiss. Estimation of optimal b-value sets for obtaining apparent diffusion coefficient free from perfusion in non-small cell lung cancer. *Physics in Medicine and Biology*, 60(20):7877–7891, 2015.

[46] Moti Freiman, Stephan D. Voss, Robert V. Mulkern, Jeannette M. Perez-Rossello, Michael J. Callahan, and Simon K. Warfield. In vivo assessment of optimal b-value range for perfusion-insensitive apparent diffusion coefficient imaging. *Medical Physics*, 39(8):4832–4839, 2012.

[47] P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

[48] Roger M. Bourne, Eleftheria Panagiotaki, Andre Bongers, Paul Sved, Geoffrey Watson, and Daniel C. Alexander. Information theoretic ranking of four models of diffusion attenuation in fresh and fixed prostate tissue ex vivo. *Magnetic Resonance in Medicine*, 72(5):1418–1426, 2014.

[49] Paddy J. Slator, Jana Hutter, Razvan V. Marinescu, Marco Palombo, Alexandra L. Young, Laurence H. Jackson, Alison Ho, Lucy C. Chappell, Mary Rutherford, Joseph V. Hajnal, and Daniel C. Alexander. InSpect: INtegrated SPECTral Component Estimation and Mapping for Multi-contrast Microstructural MRI. In Albert C. S. Chung, James C. Gee, Paul A. Yushkevich, and Siqi Bao, editors, *Information Processing in Medical Imaging*, pages 755–766, 2019.

[50] Timothy J.P. Bray, Alan Bainbridge, Naomi S. Sakai, Margaret A. Hall-Craggs, and Hui Zhang. An information-based comparison of diffusion attenuation models in normal and inflamed bone marrow. *NMR in Biomedicine*, 33(11):e4390, 2020.

[51] Gabriella Captur, Abhiyan Bhandari, Rüdiger Brühl, Bernd Ittermann, Kathryn E. Keenan, Ye Yang, Richard J. Eames, Giulia Benedetti, Camilla Torlasco, Lewis Ricketts, Redha Boubertakh, Nasri Fatih, John P. Greenwood, Leonie E. M. Paulis, Chris B. Lawton, Chiara Bucciarelli-Ducci, Hildo J. Lamb, Richard Steeds, Steve W. Leung, Colin Berry, Sinitsyn Valentin, Andrew Flett, Charlotte de Lange, Francesco DeCobelli, Magalie Viallon, Pierre Croisille, David M. Higgins, Andreas Greiser, Wenjie Pang, Christian Hamilton-Craig, Wendy E. Strugnell, Tom Dresselaers, Andrea Barison, Dana Dawson, Andrew J. Taylor, François-Pierre Mongeon, Sven Plein, Daniel Messroghli, Mouaz Al-Mallah, Stuart M. Grieve, Massimo Lombardi, Jihye Jang, Michael Salerno, Nish Chaturvedi, Peter Kellman, David A. Bluemke, Reza Nezafat, Peter Gatehouse, James C. Moon, and on behalf of the T1MES Consortium. T1 mapping performance and measurement repeatability: results from the multi-national t1 mapping standardization phantom program (t1mes). *Journal of Cardiovascular Magnetic Resonance*, 22(1):31, 2020.

[52] Marcia Morita-Sherman, Manshi Li, Boney Joseph, Clarissa Yasuda, Deborah Vegh, Brunno Machado De Campos, Marina K M Alvim, Shreya Louis, William Bingaman, Imad Najm, Stephen Jones, Xiaofeng Wang, Ingmar Blümcke, Benjamin H Brinkmann, Gregory Worrell, Fernando Cendes, and Lara Jehi. Incorporation of quantitative MRI in a model to predict temporal lobe epilepsy surgery outcome. *Brain Communications*, 3(3), 2021.

[53] Wonjeong Yang, Ji Eun Kim, Ho Cheol Choi, Mi Jung Park, Hye Young Choi, Hwa Seon Shin, Jeong Ho Won, Fei Han, Marcel Dominik Nickel, and Hyun Chin Cho. T2 mapping in gadoxetic acid-enhanced mri: utility for predicting decompensation and death in cirrhosis. *European Radiology*, 31(11):8376–8387, 2021.

[54] Uran Ferizi, Torben Schneider, Thomas Witzel, Lawrence L. Wald, Hui Zhang, Claudia A.M. Wheeler-Kingshott, and Daniel C. Alexander. White matter compartment models for in vivo diffusion MRI at 300 mT/m. *NeuroImage*, 118:468–483, 2015.

[55] Ariel Rokem, Jason D. Yeatman, Franco Pestilli, Kendrick N. Kay, Aviv Mezer, Stefan van der Walt, and Brian A. Wandell. Evaluating the Accuracy of Diffusion MRI Models in White Matter. *PLOS ONE*, 10(4):1–26, 2015.

[56] R. Mark Henkelman. Measurement of signal intensities in the presence of noise in MR images. *Medical Physics*, 12(2):232–233, 1985.

[57] Derek K. Jones and Peter J. Basser. "Squashing peanuts and smashing pumpkins?": How noise distorts diffusion-weighted MR data. *Magnetic Resonance in Medicine*, 52(5):979–993, 2004.

[58] Vladimir Golkov, Alexey Dosovitskiy, Jonathan I. Sperl, Marion I. Menzel, Michael Czisch, Philipp Sämann, Thomas Brox, and Daniel Cremers. q-Space

Deep Learning: Twelve-Fold Shorter and Model-Free Diffusion MRI Scans. *IEEE Transactions on Medical Imaging*, 35(5):1344–1351, 2016.

[59] Hanwen Liu, Qing San Xiang, Roger Tam, Adam V. Dvorak, Alex L. MacKay, Shannon H. Kolind, Anthony Traboulsee, Irene M. Vavasour, David K.B. Li, John K. Kramer, and Cornelia Laule. Myelin water imaging data analysis in less than one minute. *NeuroImage*, 210:116551, 2020.

[60] Marco Bertleff, Sebastian Domsch, Sebastian Weingärtner, Jascha Zapp, Kieran O'Brien, Markus Barth, and Lothar R. Schad. Diffusion parameter mapping with the combined intravoxel incoherent motion and kurtosis model using artificial neural networks at 3 T. *NMR in biomedicine*, 30(12), 2017.

[61] Sebastiano Barbieri, Oliver J. Gurney-Champion, Remy Klaassen, and Harriet C. Thoeny. Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI. *Magnetic Resonance in Medicine*, 83(1):312–321, jan 2020.

[62] Alberto De Luca, Alexander Leemans, Alessandra Bertoldo, Filippo Arrigoni, and Martijn Froeling. A robust deconvolution method to disentangle multiple water pools in diffusion MRI. *NMR in Biomedicine*, 31(11):e3965, 2018.

[63] R.L. Harms, F.J. Fritz, A. Tobisch, R. Goebel, and A. Roebroeck. Robust and fast nonlinear optimization of diffusion MRI microstructure models. *NeuroImage*, 155:82–96, jul 2017.

[64] Ivan I. Maximov, Farida Grinberg, and N. Jon Shah. Robust tensor estimation in diffusion tensor imaging. *Journal of Magnetic Resonance*, 213(1):136–144, 2011.

[65] Chen Ye, Daoyun Xu, Yongbin Qin, Lihui Wang, Rongpin Wang, Wuchao Li, Zixiang Kuai, and Yuemin Zhu. Accurate intravoxel incoherent motion parameter estimation using Bayesian fitting and reduced number of low b-values. *Medical Physics*, 47(9):4372–4385, 2020.

[66] Gene Young Cho, Linda Moy, Jeff L. Zhang, Steven Baete, Riccardo Lattanzi, Melanie Moccaldi, James S. Babb, Sungheon Kim, Daniel K. Sodickson, and Eric E. Sigmund. Comparison of fitting methods and b-value sampling strategies for intravoxel incoherent motion in breast cancer. *Magnetic Resonance in Medicine*, 74(4):1077–1085, 2015.

[67] Oscar Jalnefjord, Mats Andersson, Mikael Montelius, Göran Starck, Anna-Karin Elf, Viktor Johanson, Johanna Svensson, and Maria Ljungberg. Comparison of methods for estimation of the intravoxel incoherent motion (IVIM) diffusion coefficient (D) and perfusion fraction (f). *Magnetic Resonance Materials in Physics, Biology and Medicine*, 31(6):715–723, 2018.

[68] Roberta Fusco, Mario Sansone, and Antonella Petrillo. A comparison of fitting algorithms for diffusion-weighted MRI data analysis using an intravoxel incoherent motion model. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 30(2):113–120, 2017.

[69] Wenjing Chen, Juan Zhang, Dan Long, Zhenchang Wang, and Jian-Ming Zhu. Optimization of intra-voxel incoherent motion measurement in diffusion-weighted imaging of breast cancer. *Journal of Applied Clinical Medical Physics*, 18(3):191–199, 2017.

[70] Shiteng Suo, Naier Lin, He Wang, Liangbin Zhang, Rui Wang, Su Zhang, Jia Hua, and Jianrong Xu. Intravoxel incoherent motion diffusion-weighted MR imaging of breast cancer at 3.0 tesla: Comparison of different curve-fitting methods. *Journal of Magnetic Resonance Imaging*, 42(2):362–370, aug 2015.

[71] Benjamin Dallaudière, Raphaël Dautry, Pierre-Marie Preux, Anne Perozziello, Julien Lincot, Elisabeth Schouman-Claeys, and Jean-Michel Serfaty. Comparison of apparent diffusion coefficient in spondylarthritis axial active inflammatory lesions and type 1 modic changes. *European Journal of Radiology*, 83(2):366–370, 2014.

[72] Nataša Gašperšič, Igor Serša, Vladimir Jevtič, Matija Tomšič, and Sonja Praprotnik. Monitoring ankylosing spondylitis therapy by dynamic contrast-enhanced and diffusion-weighted magnetic resonance imaging. *Skeletal Radiology*, 37(2):123–131, 2008.

[73] Ying-hua Zhao, Shao-lin Li, Zai-yi Liu, Xin Chen, Xiang-cheng Zhao, Shao-yong Hu, Zhen-hua Liu, Ying-jie Mei MS, Queenie Chan, and Chang-hong Liang. Detection of Active Sacroiliitis with Ankylosing Spondylitis through Intravoxel Incoherent Motion Diffusion-Weighted MR Imaging. *European Radiology*, 25(9):2754–2763, sep 2015.

[74] Carita Tsoi, James Francis Griffith, Ryan Ka Lok Lee, Priscilla Ching Han Wong, and Lai Shan Tam. Imaging of sacroiliitis: Current status, limitations and pitfalls. *Quantitative imaging in medicine and surgery*, 9(2):318–335, 2019.

[75] D. Le Bihan, E. Breton, D. Lallemand, P. Grenier, E. Cabanis, and M. Laval-Jeantet. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology*, 161(2):401–407, 1986.

[76] Elliot R. McVeigh, Michael J. Bronskill, and R. Mark Henkelman. Optimization of MR protocols: A statistical decision analysis approach. *Magnetic Resonance in Medicine*, 6(3):314–333, mar 1988.

[77] Timothy J.P. Bray, Manil D. Chouhan, Shonit Punwani, Alanbain Bridge, and Margaret A. Hall-Craggs. Fat fraction mapping using magnetic resonance imaging: Insight into pathophysiology. *British Journal of Radiology*, 91(1089), 2018.

[78] C. M. Bishop and C. M. Roach. Fast curve fitting using neural networks. *Review of Scientific Instruments*, 63(10):4450–4456, 1992.

[79] Marco Palombo, Andrada Ianus, Michele Guerreri, Daniel Nunes, Daniel C. Alexander, Noam Shemesh, and Hui Zhang. SANDI: A compartment-based model for non-invasive apparent soma and neurite imaging by diffusion MRI. *NeuroImage*, 215:116835, 2020.

[80] Jaeyeon Yoon, Enhao Gong, Itthi Chatnuntawech, Berkin Bilgic, Jingu Lee, Woojin Jung, Jingyu Ko, Hosan Jung, Kawin Setsompop, Greg Zaharchuk, Eung Yeop Kim, John Pauly, and Jongho Lee. Quantitative susceptibility mapping using deep neural network: QSMnet. *NeuroImage*, 179:199–206, 2018.

[81] Eric Aliotta, Hamidreza Nourzadeh, and Sohil H. Patel. Extracting diffusion tensor fractional anisotropy and mean diffusivity from 3-direction DWI scans using deep learning. *Magnetic Resonance in Medicine*, 85(2):845–854, 2021.

[82] Thomas Yu, Erick Jorge Canales-Rodríguez, Marco Pizzolato, Gian Franco Piredda, Tom Hilbert, Elda Fischi-Gomez, Matthias Weigel, Muhamed Barakovic, Meritxell Bach Cuadra, Cristina Granziera, Tobias Kober, and Jean Philippe Thiran. Model-informed machine learning for multi-component T2 relaxometry. *Medical Image Analysis*, 69:101940, 2021.

[83] Noemi G. Gyori, Marco Palombo, Christopher A. Clark, Hui Zhang, and Daniel C. Alexander. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Magnetic Resonance in Medicine*, 87(2):932–947, 2022.

[84] Francesco Grussu, Marco Battiston, Marco Palombo, Torben Schneider, Claudia A.M.Gandini Wheeler-Kingshott, and Daniel C. Alexander. Deep Learning Model Fitting for Diffusion-Relaxometry: A Comparative Study. *Mathematics and Visualization*, pages 159–172, 2021.

[85] Misha P.T. Kaandorp, Sebastiano Barbieri, Remy Klaassen, Hanneke W.M. van Laarhoven, Hans Crezee, Peter T. While, Aart J. Nederveen, and Oliver J. Gurney-Champion. Improved unsupervised physics-informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients. *Magnetic Resonance in Medicine*, 86(4):2250–2265, 2021.

[86] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[87] Elisa Scalco, Giovanna Rizzo, and Alfonso Mastropietro. The quantification of intravoxel incoherent motion – mri maps cannot preserve texture information: An evaluation based on simulated and in-vivo images. *Computers in Biology and Medicine*, 154:106495, 2023.

[88] Serge Didenko Vasylechko, Simon K. Warfield, Onur Afacan, and Sila Kurugol. Self-supervised ivim dwi parameter estimation with a physics based forward model. *Magnetic Resonance in Medicine*, 87(2):904–914, 2022.

[89] Djork Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2015.

[90] Sean C. Epstein, Timothy J.P. Bray, Margaret A. Hall-Craggs, and Hui Zhang. Task-driven assessment of experimental designs in diffusion MRI: A computational framework. *PLOS ONE*, 16(10):e0258442, oct 2021.

[91] Zhenghan Fang, Yong Chen, Weili Lin, and Dinggang Shen. Quantification of relaxation times in MR Fingerprinting using deep learning. *Proceedings of the International Society for Magnetic Resonance in Medicine ... Scientific Meeting and Exhibition. International Society for Magnetic Resonance in Medicine. Scientific Meeting and Exhibition*, 25, 2017.

[92] Cagdas Ulas, Dhritiman Das, Michael J. Thrippleton, Maria del C. Valdés Hernández, Paul A. Armitage, Stephen D. Makin, Joanna M. Wardlaw, and Bjoern H. Menze. Convolutional Neural Networks for Direct Inference of Pharmacokinetic Parameters: Application to Stroke Dynamic Contrast-Enhanced MRI. *Frontiers in Neurology*, 9:1147, 2019.

[93] Simin Li, Jian Wu, Lingceng Ma, Shuhui Cai, and Congbo Cai. A simultaneous multi-slice t2 mapping framework based on overlapping-echo detachment planar imaging and deep learning reconstruction. *Magnetic Resonance in Medicine*, 87(5):2239–2253, 2022.

[94] Erich L. Lehmann and George Casella. *Theory of Point Estimation.* Springer-Verlag, New York, NY, USA, second edition, 1998.

[95] David J. C. MacKay. *Information theory, inference, and learning algorithms.* Cambridge University Press, 2003.

[96] Marian A. Troelstra, Anne-Marieke Van Dijk, Julia J. Witjes, Anne Linde Mak, Diona Zwirs, Jurgen H. Runge, Joanne Verheij, Ulrich H. Beuers, Max Nieuwdorp, Adriaan G. Holleboom, Aart J. Nederveen, and Oliver J. Gurney-Champion. Self-supervised neural network improves tri-exponential intravoxel incoherent motion model fitting compared to least-squares fitting in non-alcoholic fatty liver disease. *Frontiers in Physiology*, 13, 2022.

[97] Chen Ye, Daoyun Xu, Yongbin Qin, Lihui Wang, Rongpin Wang, Wuchao Li, Zixiang Kuai, and Yuemin Zhu. Estimation of intravoxel incoherent motion parameters using low b-values. *PLOS ONE*, 14(2):1–16, 2019.

[98] Eric Aliotta, Hamidreza Nourzadeh, Jason Sanders, Donald Muller, and Daniel B. Ennis. Highly accelerated, model-free diffusion tensor mri reconstruction using neural networks. *Medical Physics*, 46(4):1581–1591, 2019.

[99] Sean C. Epstein, Timothy J. P. Bray, Margaret Hall-Craggs, and Hui Zhang. Choice of training label matters: how to best use deep learning for quantitative MRI parameter estimation. 2022.

[100] Davood Karimi, Camilo Jaimes, Fedel Machado-Rivas, Lana Vasung, Shadab Khan, Simon K. Warfield, and Ali Gholipour. Deep learning-based parameter estimation in fetal diffusion-weighted mri. *NeuroImage*, 243:118482, 2021.

[101] Gemma L. Nedjati-Gilani, Torben Schneider, Matt G. Hall, Niamh Cawley, Ioana Hill, Olga Ciccarelli, Ivana Drobnjak, Claudia A.M. Gandini Wheeler-Kingshott, and Daniel C. Alexander. Machine learning based compartment models with permeability for white matter microstructure imaging. *NeuroImage*, 150:119–135, 2017.

[102] Oliver J. Gurney-Champion, Remy Klaassen, Martijn Froeling, Sebastiano Barbieri, Jaap Stoker, Marc R. W. Engelbrecht, Johanna W. Wilmink, Marc G. Besselink, Arjan Bel, Hanneke W. M. van Laarhoven, and Aart J. Nederveen. Comparison of six fit algorithms for the intra-voxel incoherent motion model of diffusion-weighted magnetic resonance imaging data of pancreatic cancer patients. *PLOS ONE*, 13(4):1–18, 04 2018.

[103] João P. de Almeida Martins, Markus Nilsson, Björn Lampinen, Marco Palombo, Peter T. While, Carl-Fredrik Westin, and Filip Szczepankiewicz. Neural networks for parameter estimation in microstructural mri: Application to a diffusion-relaxation model of white matter. *NeuroImage*, 244:118601, 2021.