Validation in the age of machine learning: A framework for describing validation with examples in transcranial magnetic stimulation and deep brain stimulation

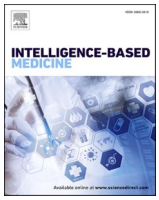
J.S.H. Baxter, P. Jannin

Intelligence-Based Medicine 7 (2023) 100090

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine



Validation in the age of machine learning: A framework for describing validation with examples in transcranial magnetic stimulation and deep brain stimulation

John S.H. Baxter *, Pierre Jannin

Laboratoire Traitement du Signal et de l'Image, INSERM UMR 1099, Université de Rennes 1, Rennes, France

ARTICLE INFO

Keywords:
Validation
Evaluation
Bias
Medical information processing
Ground truth
Transcranial magnetic stimulation
Deep brain stimulation

ABSTRACT

Medical information processing is a staple of modern medicine with its increasing focus on the collection of numeric medical data such as questionnaires, biophysiological signals, and medical images. Although these modalities have long existed and guided medical practice, the movement towards using algorithms to transform, curate, summarise, and otherwise interact with this data is relatively new. Novel algorithms now form the interface between clinical users and data, extracting information that would otherwise be inaccessible or cumbersome. Recently, machine learning has expanded the capacities of these algorithms, using *a priori* acquired (and often annotated) datasets to learn a complex computational task. Validation of these techniques is inherently important for determining their safety and efficacy in a particular clinical context. However, methodological considerations such as the definition of reference data and validation procedures can obscure validation issues such as inaccurate reporting, a lack of standardisation, and a variety of biases. The purpose of this paper is to develop a framework for understanding medical information processing algorithms with a focus on validation that is adapted for machine learning approaches as well as traditional ones. This framework is instantiated in two example literature reviews which serve as the starting point for a discussion on how validation can be improved cognisant of machine learning.

1. Introduction

Validation is undeniably critical to the use of medical data processing algorithms both in research and in clinic. For the latter, various levels of validation are not only desirable, but are mandated through regulatory bodies in order to ensure patient safety and treatment efficacy [25,27,76]. Although medical information/image processing/computing (MIC) algorithms were formerly more reliant on human interpretation of a small number of modalities (e.g. recording observations about the patient's symptoms, visual examination of x-ray images, or auditory examination using a stethoscope), there is a growing number of new modalities (smartphone cameras, gyroscopes, etc …) and representations derived from existing modalities (i.e. data processing) that assist medical decision making [61]. Furthermore, the rise of machine learning algorithms in computer vision has provided medical signal and image processing research in particular with a large number of new and highly performant tools which have yet to be fully integrated into the conceptual framework of MIC algorithm validation which has been developed with traditional algorithms in mind [76].

There has been an effort in the community towards standardised tools, allowing for a higher degree of reproducibility owing to the use of dedicated software libraries such as ITK [44], MeVisLab [21], and Slicer [29,31]. These tools make the reproduction of methods easier (given the underlying code) by delivering a software environment that is higher-level and more independent of the exact computational infrastructure. However, the ability to reproduce a paper's code exactly does not necessarily mean that the method itself is reproducible, especially given the diverse array of validation procedures and techniques used in the literature. This is critical as scientific growth and clinical utility rely on independent reproduction, which implies that at least some elements of the validation process, not only the data used, have been changed while still getting the predicted results. Especially with algorithms relying on trained models, a researcher reproducing a method from the literature might be unsure if a deviation from the reported results comes from a possible error in its re-implementation, a critical difference in clinical context, or simply a difference of validation. Many aspects of

* Corresponding author.
E-mail address: jbaxter@univ-rennes1.fr (J.S.H. Baxter).

https://doi.org/10.1016/j.ibmed.2023.100090
Received 22 April 2022; Received in revised form 22 November 2022; Accepted 4 January 2023
Available online 11 January 2023

image pre-processing in particular go largely unnoted even when they can have a large effect on the result. For example, many papers that use machine learning to analyse CT images neglect to give the parameters of the kernel used in their reconstruction, which can sometimes result in differences on the same order of magnitude as what they see between modern state-of-the-art methods [7].

There are a number of immediate pitfalls one must overcome in any review covering a topic as essential as validation in an area as broad as MIC itself. The first is that many of the techniques used come from a wide variety of fields: software engineering and computer science, computational statistics and machine learning, psychology and psychometrics, etc…For example, the word *validation* in software engineering means *assessing whether or not a particular system fulfils the purpose to which it was intended* and is contrasted with terms such as *verification* [79]. In machine learning, on the other hand, *validation* means *assessing how performant a system is on previously unseen data* [75]. Even within an individual discipline, particular words can take on a variety of meanings. In machine learning, *validation data* sometimes means the unseen data mentioned above, but in other contexts may mean a particular fraction of the whole dataset that is used to determine hyper-parameters, making it conceptually similar to 'Training' data. Thus, any framework concerning validation must clearly define terms and remove as much ambiguity as possible.

For the purposes of this paper, we will be focusing on validation as the process of determining whether or not a system ***fulfils its intended clinical purpose*** as opposed to whether or not it is correctly implemented. We do so in a constructive manner, emphasising the main steps used in the construction and validation of MIC algorithms as outlined in Fig. 1 This model is both an extension and a distillation of the earlier approach taken by Jannin et al. [26] and the checklist proposed by Maier-Hein et al. [41] adapted for the nuances arising from new machine learning approaches while showing commonalities with traditional approaches. This paper is constructed as follows.

- We will propose a framework for describing the general process of validating a MIC algorithm including the flow of both data and decisions. This allows us to identify common features of validation procedures across a wide range of contexts and develop a quasi-exhaustive list of validation details that should be reported in any given study.
- We will apply this framework to a literature review of machine learning algorithms in cortical point localisation for transcranial magnetic stimulation and for subthalamic nucleus segmentation for deep brain stimulation in order to draw conclusions about how validation is performed in these particular sub-fields, analysing common correctable mistakes that lead to large biases in the validation results.
- Lastly, we will use this knowledge of common issues, as well as our framework to produce concrete, high-level recommendations for improving MIC algorithm validation.

The intent is to motivate the research community to build awareness and encourage more robust validation procedures. The benefit of this is two-fold: it may smooth the transition of their research into clinical use and it may also improve our capability as a community to draw conclusions about novel methods across a variety of published results. This is because it allows us to better find papers that are comparable not only in terms of technical details, but also validation details that can also strongly affect quantitative results.

2. Framework for the validation of MIC algorithms

The overall structure of our framework is presented in Fig. 1 with its specific terms defined in Table 1, which is designed to mimick the process of performing an idealised experiment in validating a MIC algorithm and can, at a high-level, be read in a left-to-right movement. In practice, this is often not equivalent to the order in which the precise computational steps are performed, but can be thought of more as a type of loose dependency. This diagram is designed more to handle the semantic dependencies between different decisions and processes which should be considered prior to actually performing validation and should be reported in any resulting publications. Symbols have been given to each of the components to allow for a readily-used short-hand.

Boxes in this image represent a collection of decisions to be made or information to be reported that describe and structure the validation procedure with arrows indicating dependency with the variables in one box relying on knowledge of the other. Naturally, some of these aspects may be unnecessary for certain algorithms or validation procedures (e.g. if only a single model is ever evaluated, there is no need for to perform aggregation or selection to extract a resultant model) but this is for the sake of creating a sufficiently general framework to capture the essence of different validation procedures across technical methodologies and clinical contexts.

As such, this model is an extension of the one created by Jannin et al. [26] which maintains a similar flow structure which they used later in an image-guided interventions specific context [27]. However, this paper extends outside of their scope, allowing for emerging technical methods (such as machine learning) to be readily expressed for a broad range of applications and integrated with more traditional methods. For each part of the framework, an example will be given from

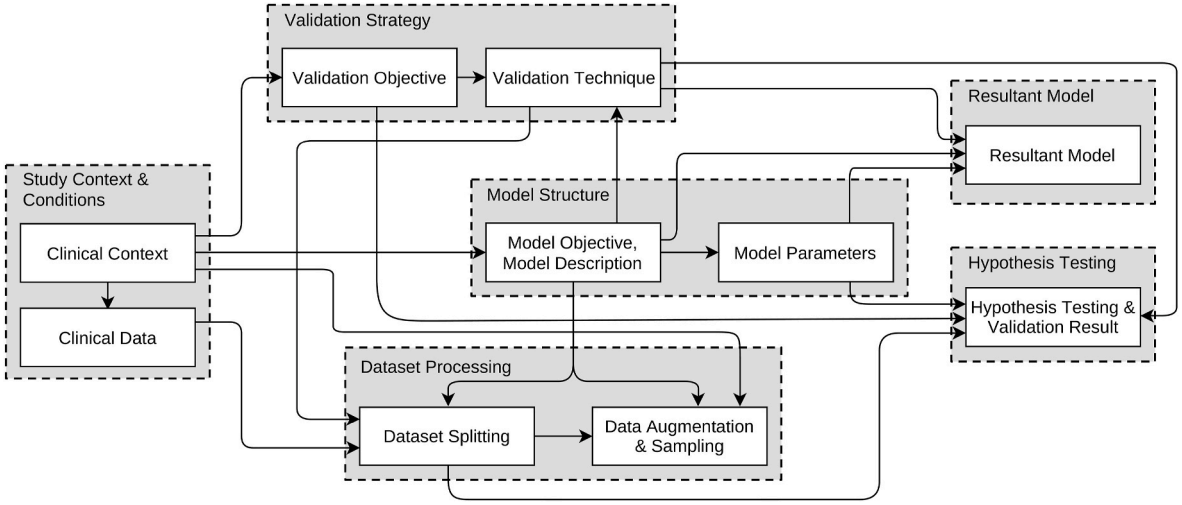


Fig. 1. The main components of the validation procedure for MIC algorithms in terms of the decisions made and their dependencies.

Table 1
Summary of validation characteristics to be reported and corresponding symbols.

	Brief Description
Clinical Context	
C	The general context around the medical procedure being considered
L_A	The level of assessment
L_I	The level of interaction
T_H	The task being performed by the clinician using the algorithm
T_I	The information provided by the clinician for the algorithm
T_A	The specific task of the algorithm
Clinical Data	
P_P	The patient distribution being used, common features, gender, diagnosis, etc …
P_D	The distribution of the data (i.e. technical characteristics such as modality, type, resolution, etc …)
P_I	How the data is pre-processed and/or post-processed (e.g. reconstruction, normalisation …)
S_{ID}	The specifics of the input data format
S_{OD}	The specifics of the output data format
A	Assumptions made about the data or processing
Model Structure	
G_F	The granularity (e.g. slice, 2D/3D patch, volumetric image …) used for training and application
F_O	The objective of the model including any pre-conditions, any post-conditions, etc …
F_D	Thefixed parameters and structure of the model not subject to change (e.g. type of model, architecture, loss, optimiser, initialisation, etc …)
F_P	The parameters of the model that are subject to change (e.g. weights, trainable hyperparameters, etc …)
Validation Objective	
V	The specific criteria or aspect under evaluation
M_V	The validation metrics used as a surrogate for these criteria
I	What reference is used and how is it collected
M_Q	The qualifying performance, the threshold for the metrics that allow the method to be clinically useable.
M_{OU}	The observation uncertainty or the measure of the variability in equivalent data.
M_{RU}	The reference uncertainty or the measure of the variability in valid references.
Validation Technique	
V_T	The specific validation technique used.
G_V	The validation granularity (i.e. patient, centre, etc …).
Dataset Splitting	
N_{test}	The number of data points used to determine M_O.
N_{train}	The number of data points used to determine F_P.
V_R	The number of times the experiment is repeated.
Data Augmentation & Sampling	
A_{IE}	In/equi-variances, their range, and importance
A_B	Potential biases arising from data augmentation
F_A	The type and extent of data augmentation applied as well as to what data.
F_S	The sampling method used including any modifications (e.g. adaptive sampling, class balancing)
Hpyothesis Testing & Validation Result	
M_O	The distribution of observed validation results
F_H	The method used to verify the hypothesis
Resultant Model	
F_R	How the resultant/final model is constructed.

ultrasound-guided, MRI-planned prostate cancer needle biopsy involving image registration [69].

2.1. Clinical context

The first aspect of validation is to elucidate the context and conditions in which it takes place, considering it both from the point of view of the clinical workflow and also the available data (see Fig. 2). This can be understood as the first half of the *assessment phase* suggested by Jannin et al. [27] in which the underlying clinical problem and resources are elucidated.

The *clinical context* is the particular workflow into which this MIC algorithm is to be integrated as well as what role the algorithm is to have. Often, clinical systems are not composed of a single data processing step, but rather a number of interacting components each with a particular goal and particular pre-conditions. The goal of stating the clinical context in a clear manner is to elucidate considerations and information for later decisions, ensuring that clinical utility is always kept in mind. Medical imaging systems are becoming increasingly modular, so it is therefore of increasing importance that the individual parts can be validated separately without losing sight of how they interact with others. In order to clarify the context for the particular MIC algorithm, it is important to clearly state.

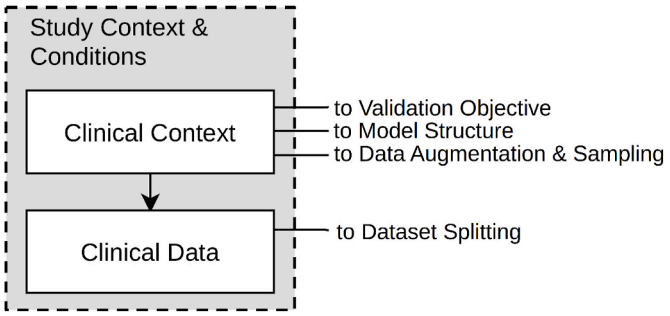


Fig. 2. Study Context & Conditions consisting of the Clinical Context which describes the general workflow in which the algorithm will eventually be situated and the Clinical data which describes the characteristics of the dataset used in the validation procedure.

- C, the broad clinical context (e.g. pathology being considered, clinical workflow, type of intervention …);
- L_A, the *level of assessment* which describes in the most general terms the type of validation being performed;
- L_I, the *level of interaction* between the clinical user and the algorithm;
- T_H, the *human tasks* being performed by the clinician that most immediately contains the image processing algorithm under consideration;
- T_I, the *interaction tasks* performed by the clinician solely to communicate information to the algorithm, as well as what mechanisms are provided for said tasks; and
- T_A, the specific *algorithmic tasks* being performed by the image processing algorithm (i.e. segmentation, registration, fusion, etc …).

(An example of an instantiation of these variables is given in Table 2.)

Outside of the general clinical context of the algorithm, it is important to have a high-level representation of the context of the validation procedure, that is, the *level of assessment*, L_A, being performed. Fryback & Thornbury [14] described a series of L_A's (later expanded upon by Jannin et al. [26]) which can be interpreted as a spectrum from the technical to the societal.

Table 2
Example clinical context from Ref. [69]. Quotations are taken directly from the text. Non-quotations are either implicit in the text or are generated for the sake of providing an example.

Clinical Context	
C	"In this regard, MR-TRUS registration technique provides an effective way to use TRUS to target biopsy needles toward regions of the prostate containing MR identified suspicious lesions "
L_A	Medium-low level - "feasibility on clinical data"
L_I	Semi-automatic
T_H	"… transrectal ultrasound (TRUS) guided prostate biopsy is the standard approach for definitive diagnosis and guiding biopsy needles to suspicious regions in the prostate …"
T_A	Iteratively refined MRI to US registration - "In this work, we propose a novel duality-based approach to computing the challenging 3D MIND-based non-rigid MR-TRUS deformable registration."
T_I	"We initialize the registration using 3 manually placed approximately corresponding landmarks"

- **low-level** measurement of a specific technical property in a highly controlled setting (i.e. accuracy on phantom data, computation time, resolution, memory, etc …);
- **medium-low-level** measurement of a property which is necessarily conditioned on human data (i.e. performance on human data, ease-of-use, population coverage, etc …);
- **medium-level** measurement of *in situ* clinical performance (i.e. improvement in real workflow efficiency, margin of error in clinic, acceptability to clinicians, immediate effects on symptoms or physiology, etc …);
- **medium-high level** measurement of clinical response (i.e. improved patient outcomes, morbidity, etc …); and
- **high-level** validation of how well the proposed system fits into the healthcare ecology (i.e. public health evaluation, economic evaluation, ethics, quality-of-life, etc …).

The L_A is important to determining what technical parameters (e.g. the degree of completeness of the algorithm from prototype to robust clinic-ready application) and validation parameters (e.g. metrics, mitigation of different biases through dataset splitting …) are applicable. Although often a single L_A is used, sometimes two consecutive ones are addressed, especially in papers with multiple sub-experiments. Often, higher L_A's implicitly assume that lower ones have already been investigated and the algorithm's more basic capabilities confirmed.

The goal of separating the human and algorithmic tasks is to clarify exactly where the algorithm fits into the clinical workflow in relation to the larger task being performed by the human operator. The *level of interaction*, L_I, reflects that some advanced algorithms are designed to be interactive, receiving information from the clinical user to guide and improve quality. Example of higher L_I's can be found in visualisation algorithms in which the user must specify which aspects of the data are to be visualised, but they also exist in data processing algorithms such as image segmentation [2]. The L_I ranges between.

- **manual** when the task is performed solely by the user;
- **interpolated** when the task is performed largely by the user for particular parts of the data and the role of the computer is to interpolate the remainder;
- **interactive** when the user iteratively provides information to the computer which it then learns or extrapolates from to perform the algorithmic task;
- **semi-automatic** when the algorithm requires a well-defined and highly limited amount of information from the user, likely in the form of input initialisation or output selection; and
- **fully automatic** when there is little to no user input.

A higher L_I can complicate validation, especially at a lower L_A, as it introduces uncertainty into the experiment which should be measured and reported.

The L_I informs how the immediate clinical task is decomposed into tasks that are performed by the human versus those performed by the computer. The *human tasks* (T_H) are those being performed by the clinician, which is in many contexts either directly or indirectly interacting with the patient. This could include diagnostic tasks, such as determining whether or not a given patient has a particular condition, or physical tasks such as moving a surgical instrument to a particular area. The key element of the T_H is that it can make the clinical context more precise. For example, the clinical context may be the planning of some particular surgery, with the T_H specifically being the visualisation of the implicated anatomy.

Unlike the T_H which relates the clinician's role to the wider clinical context, the *interaction task* (T_I) defines the role that the clinician takes as a user of the MIC system. For fully automatic systems, this task may be as simple as specifying what data to process, but in other cases, it could be more complex and structured [2]. For example, the T_H may be to visualise some anatomy, and the T_I is the user clicking on the particular anatomy to be visualised. The *algorithmic task* (T_A) is the high level processing performed by computer using the information provided by through the T_I. Note that the T_I and T_A may be iterative with both the clinician and the computer repeatedly exchanging information (e.g. corrections, annotations, changes in parameters …) which can complicate their measurement [2].

There are a couple of ways in which these iterative processes can be validated. The first is through simulating the users themselves either computationally [48,83,84] or mathematically [80]. Whether or not this should be considered *good* validation then largely depends on whether the simulation reaches a level of realism appropriate to the level of assessment, with more general or less-realistic simulated users tending towards the lower end of the spectrum. An alternative is to perform a user study. However, the question of their quality then comes down to exactly what is being measured (i.e. accuracy or usability) as well as the reference used for these metrics [4,71].

2.2. Clinical data

For any given experiment, the clinical context has an influence on the available *clinical data*. At a surface level, the clinical data for a MIC task are the images/information themselves which are to be inputted into the algorithm, although there may be a large amount of supporting data such as clinical questionnaires, patient history, surgical instrument tracking, etc…to support the validation process. In diagnostic systems, this data may be crucial for contextualising the process as a whole, determining which patients (or disease sub-types) are present at the different stages of the diagnostic workflow. For interventional systems, the clinical data can help determine what alternative approaches are feasible, including back-up approaches in the case of (partial or total) system failure. The data being used in the validation of the algorithm can be broadly summarised by.

- P_P, the distribution of the patients and healthy controls used including key demographic information such as age, gender, diagnosis, and relevant clinical information,
- P_D, the characteristics of the data being used, specifically the properties and characteristics of the medical imaging and other supporting data used as well as their intrinsic characteristics (e.g. modality, spatial/temporal/intensity resolution, tissue contrast),
- P_I, the pre- or post-processing that the data has gone through outside of the scope of the model itself. For medical images, this often includes common steps such as image reconstruction or bias field correction, but may also include manual steps,
- S_I and S_O, the data input and output formats respectively to the algorithm including file type, image orientation, sampling rate/resolution, etc …
- A, the set of assumptions about the data used to simplify the problem.

(An example of an instantiation of these variables is given in Table 3.)

Often, the most informative aspect of the data from the clinical point-of-view is the patient distribution, P_P. It has been long known that demographic information, sex and gender in particular, have a strong effect on the characteristics of a wide array of data modalities. There are a large body of literature on gender differences and their effect on particular anatomy that may not be immediately expected by an algorithm designer such as cartilage density [10], bone tissue characteristics [51] and a plethora of neural phenomena [37,81]. The degree to which we understand systematic differences like these is constantly changing and thus reporting as much as possible may allow for biases to be elucidated later based on knowledge that was unavailable to the algorithm designers during the validation procedure itself. Although the amount and type of necessary information varies widely from application to application, it is also unfortunately frequent for this information to be largely missing as shown even in the illustrative example in

Table 3
Example clinical data from Ref. [69]. Quotations are taken directly from the text. Non-quotations are either implicit in the text or are generated for the sake of providing a reasonable example.

Clinical Data	
P_P	10 patients (M = 10) - "We performed the proposed method to register 10 patient images." No additional information given.
P_D	"In this study, T2-weighted MR images using a body coil and corresponding 3D TRUS images from 10 patients were acquired. The MR images were obtained at 3 T using a GE Excite HD MRI system (Milwaukee, WI, USA) at an image size of 512 × 512 × 36 voxels with a voxel size of 0.27 × 0.27 × 2.2 mm3. The 3D TRUS images were acquired using a 3D TRUS mechanical scanning system developed in our laboratory, using a Philips HDI-5000 US machine with a Philips C9-5 transducer. The 3D TRUS image size is 448 × 448 × 350 voxels with a voxel size of 0.19 × 0.19 × 0.19 mm3."
P_I	"The MR image is first resampled to have the same dimensions and voxel size as the TRUS image. We initialize the registration using 3 manually placed approximately corresponding landmarks and the centroid of the three points as a default point on the 3D TRUS and MR images to generate a rigid transform as initial alignment."
S_{ID}	Not specified - possibly "MR images are saved in DICOM format in RAS orientation, 3D TRUS images are saved in DICOM format"
S_{OD}	Not specified - possibly "3D deformation image saved in DICOM format"
A	Not specified - possibly "3D TRUS reconstruction is performed successfully"

Related to P_P is the data characteristics (P_D). For clinical datasets in which an exact P_D cannot be reasonably specified in its entirety, giving its parameters can indicate the scope of data on which the proposed method can be applied. At the most basic level, this would be simply identifying the modalities used (i.e. CT, MRI, EEG, etc …) with enough technical information (imaging parameters, sampling rates, number of images/channels, etc …) to allow an equivalent dataset to be collected. From a reproducibility perspective, including the specific data format (S_I and S_O) can largely improve the chances of a paper's methods being successfully ported to another centre and minimize a number of potential technical difficulties in repeated independent validation. Dataset pre-processing or post-processing, P_I, is often used to make the distribution of the data seen or produced by the algorithm more homogeneous, but itself is subject to error and uncertainty and thus warrants being reported. Certain forms of pre- and post-processing are known to be more stable or repeatable than others, but given their ubiquity, are rarely validated as a component of the overall model under consideration. In these cases, attention should still be taken either to validating the appropriateness of the additional processing or an explicit citation to another paper with the focus on doing exactly that for a similar or equivalent patient population. As mentioned in the Introduction, even implicit pre- and post-processing can have an effect on performance and thus being explicit about these processes is crucial for reproducibility.

At the lowest L_A, simulated or phantom data is often used as a surrogate for patient data meaning that, P_P is not truly available, although P_D can be specified exactly rather than heuristically. This simulated data could be in the form of physical objects with known characteristics (i.e. phantoms) or numeric simulations of physically well-understood phenomena. Regardless of whether the simulation is physical, numeric, or a combination of the two, there is always a trade-off between simulation fidelity and availability, that is, that simulations can never simultaneously mimic an actual patient closely and remain easy to acquire in large quantities [17]. Thus, the use of simulated data introduces its own set of biases to the experiment, favouring algorithms that rely on heavier assumptions about the data (e.g. lack of movement, uniform contrast, etc …) that are not reflective of clinical reality [66].

Lastly, any algorithm designer must make assumptions, A, about the data that render the problem more feasible. These assumptions can be innocuous, e.g. the particular anatomy of interest will be present or be in a particular range of sizes, or that there will be an absence of a particular pathology that would render the problem more difficult. Additionally, some assumptions may be made about how the data is annotated which can have a strong effect on the quality of the reference data used. Thus, these assumptions introduce different levels of bias into the validation study that can be measured, reported, critiqued, and refined. One of the more difficult assumptions to catch concern the distribution of the data itself, as we often implicitly assume that the dataset collected is representative of the clinical situation, which could be false due to underlying racial [70], gender [23,35], or age [70] biases amongst other sources.

2.3. Model structure

After the clinical context and clinical data have been determined, it is now possible to design the particular *model* (i.e. algorithm, process, machine learning architecture, etc …) to address the computational task (as per Fig. 3). The model structure forms a central role in a validation study, which is to be expected as it is the element that is under the more direct investigation. Although even a cursory review of the plethora of models used in almost any general MIC task is far beyond the scope of this paper, it is important for all of them to distinguish between four basic elements, the *training/application granularity*, the *model objective*, the *model description*, and the *model parameters*.

Granularity is defined as the amount of data is considered as single data-point for various purposes. Thus, the training and application granularities, G_F, may not be a single value but multiple - one part of the model may use a coarser or a finer G_F compared to another. For example, in MIC, entire volumetric images may be acquired and pre-processed, but the machine learning component processes them in a slice-by-slice manner to conserve memory, meaning that G_F differs between acquisition/pre-processing and processing.

Briefly, the *model objective*, F_O, is the computation that the model is trying to perform or the particular post-conditions it is trying to fulfil. This may sometimes be indirectly related to the T_A, especially in machine learning systems in which F_O is more often the optimisation of a particular loss function which determines elements of the model or performs the desired algorithmic task as a by-product of this optimisation. Outside of the context of optimisation, F_O can often be thought of in simpler terms, abstracting the computation performed by the model and specifying at a high level the algorithms pre- and post-conditions.

The *model description*, F_D, is the collection of design decisions that are

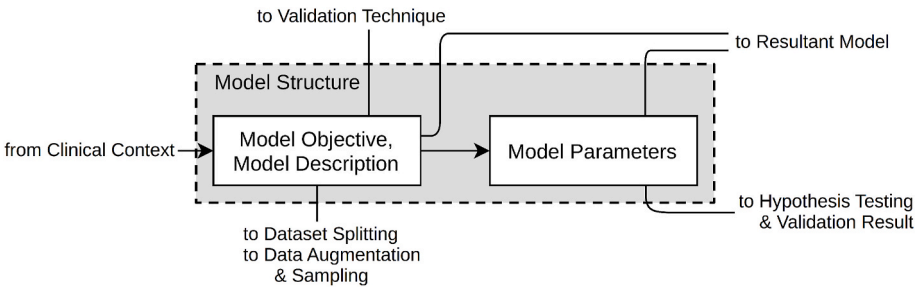


Fig. 3. Model Structure consisting of the Model Objective which the specific technical task or goal the model is designed towards, the Model Description which describes the fixed elements of the model, and the Model Parameters which describes the elements of the model which are adjusted to fit to the data.

fixed prior to the algorithm having access to any of the data. These can include aspects such at the type of model being used (e.g. a neural network), its architecture (e.g. number of layers and their widths), as well as aspects that control the model creation process, such as the optimiser used. Some parameters, such as learning rates, regularisation weights, etc…are set ahead of time or are learned through a process often called hyperparameter tuning. In the latter case, it is important to clearly state so as these hyperparameters are technically learnt and thus do not fall into this category. The rule-of-thumb is that F_D contains everything that should be affected by the general aspects of the data (P_P, P_D, etc …), but not the particular data used. Unlike F_O, F_D is configurable, that is, different F_D's could be used to perform the same task. Instead of being inherent to the problem being solved, F_D elucidates design decisions that could be modified or ablated to create similar models. For an example, consider neural network architectures: it is easy to image two neural networks with a different number and configuration of layers that nevertheless could be trained on the same data using the same loss function. This is frequently used as the basis for an ablation study in papers that propose new architectures [1,4,47].

In contrast, the *model parameters*, F_P, which set as a function of the computational process investigating a subset of the data, i.e. they are learned. This is specifically done using only 'Training' data, the definition of which will be discussed in a subsequent section. For the purposes of preventing *data leakage* it is crucial to identify which elements of the model fall into which category and to maintain a clear separation between the two in order to prevent biasing the validation results [30]. There is a possibility that a technique does not involve any F_P, e.g. the hyperparameters and weights are inferred from an understanding of the problem domain or separate simulated data rather than learnt from the clinical data. In the context of machine learning, this is rare as even models that are designed to learn on-the-fly still involve hyperparameters which have likely been set cognisant of the available data.

The correct divide between F_D and F_P can also depend on the level of assessment. For example, at higher levels of assessment, such as in a prospective clinical trial, there is a strong possibility that the model has already been trained and the parameters set in a previous publication (such as how [65] performs a clinical trial to evaluate their method published as [9]). In these cases, there is a distinct benefit to mentioning that the model is already trained, simply to differentiate it from papers in which a model is retrospectively applied (i.e. lower levels of assessment) as merely mentioning that a model is machine-learning-based would give the casual reader the impression that the learner process is on-going rather than completed.

Managing this divide is particularly important not only for prevent data leakage (discussed in Section 2.5) but also for managing more subtle aspects of methodology-centric bias; which has been previously elaborated upon in Ref. [3].

2.4. Validation objective

The validation objective generally describes what aspects of the algorithm are being validated ranging from an abstract description of those aspects to the concrete way in which they are measured. (See Fig. 4 for the combined placement of the Validation strategy as a whole.) The elements of the validation objective are.

- V, the validation criteria under consideration (i.e. robustness, feasibility, efficacy, etc …)
- M_V, the set of metrics being used to measure V,
- I, the information extracted from the data that determines the *reference* (or *ground truth*) and how it is extracted,
- M_Q, the *qualifying performance*, i.e. the worst values of M_V that would be still considered clinically useable or a meaningful technological improvement, and, if possible:
- M_{OU}, the *observation uncertainty*, i.e. the range of M_V attributed to noise or other irreducible features outside of the control of the evaluated algorithm, which contextualises small variations or differences in the quantitative results in terms of potential mitigating factors,
- M_{RU}, the *reference uncertainty*, i.e. the range of M_V applied to I instantiated under different conditions, for example, measures such as *inter-rater* and *intra-rater* variability or the noise characteristics.

(An example of an instantiation of these variables is given in Table 4.) The purpose of the first four is to ensure that the work is clinically useful and reproducible whilst the last two contextualise the quality and ideal range of the validation objective.

The purpose of the validation objective, V, is to define what property of the system is being validated. In essence, V is a concrete instantiation of L_A. These properties should be phrased as unambiguously beneficial for a system to have (e.g. accuracy, precision, robustness, consistency, speed, error tolerance, etc …) or unambiguously detrimental to have (e.g. cost, energy consumption, memory consumption, etc …) [25]. The exact meaning of these objectives can be nuanced and subject to

Table 4
Example validation objective from Ref. [69]. Quotations are taken directly from the text. Non-quotations are either implicit in the text or are generated for the sake of providing a reasonable example.

Validation Objective	
V	Accuracy, speed
M_V	"We measured the target registration error (TRE) as the overall misalignment of manually marked corresponding intrinsic fiducials in MR and 3D TRUS images. […] We also measured the fiducial localisation error (FLE) to allow determination whether fiducial identification dominates the TRE. We also compared the registered MR and corresponding 3D TRUS images by calculating the Dice similarity coefficient (DSC), the mean absolute surface distance (MAD), and the maximum absolute surface distance (MAXD). All validation metrics were separately calculated for three prostate sub-regions: the apex, mid-gland and base, selected along the apex-base axis of the manual segmented TRUS prostates (0.3, 0.4, 0.3 of the length of the base-apex axis respectively)."
I	"We selected 41 fiducial pairs, of which 17 were within the peripheral zone (PZ), in which up to 80% of the tumors can be located."
M_Q	"clinically acceptable maximum TRE of 2.5 mm"
M_{OU}	Not specified - possible sources could include image noise, 3D TRUS reconstruction artefacts due to motion, etc …
M_{RU}	"We also measured the fiducial localisation error (FLE) to allow determination whether fiducial identification dominates the TRE …. FLE are 0.21 mm for 3D TRUS and 0.18 mm for MR."

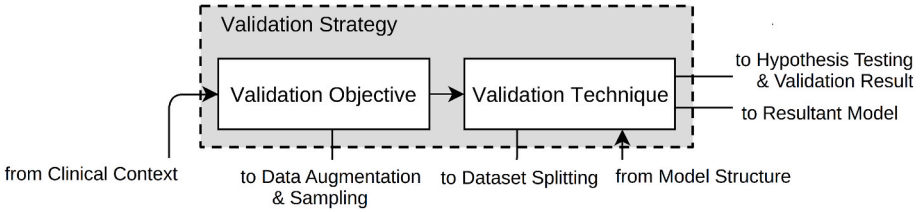


Fig. 4. Validation Strategy consisting of the Validation Objective which defines the particular aspect being validated and how it is to be measured, and the Validation Technique which determines how the data will be partitioned, how the model(s) will be constructed, and how the results will be aggregated.

differences across domains and algorithm types. For example, if a medium-low L_A is desired (i.e. validating the technical feasibility of the algorithm on clinical data) that could mean different things in different contexts. In a diagnostic context, it could mean that high accuracy with constraints on time and memory, whereas the opposite may apply for a surgical system where the fastest system meeting a minimum accuracy would be preferred. As a system matures (i.e. moves from the research environment into clinical use) and the L_A gets higher, the choice of V also changes to reflect the evolved purpose of the validation.

At this stage, V is still somewhat abstract and must be made concrete in order to be useable. Specifically, it is substituted by a set of metrics, M_V, whose measurements can be used as a surrogate or definition for the desired quality. These sometimes involve comparison against a *reference*, I, which we define as *an ideal feasible output that is defined outside of the algorithm itself, i.e. via a more accurate modality, in-depth physical understanding, manual processing, or simulation*. Aside from being representative of the clinical quality defined by the clinical context, in order for M_V to be useful and reproducible, it must have some particular characteristics.

- *practicality,* allowing them to be determined with minimal intrusion into clinical workflow or patient care;
- *quantifiability,* allowing for easier comparison;
- *monotonicity,* meaning that any positive or negative trend in the metric can be readily interpreted as reflecting positively or negatively on the algorithm's quality;
- *aggregability,* meaning that the results derived from multiple raters/runs/trials/etc…can be combined, summarised, and interpreted statistically.

The last three characteristics are listed in the order of increasing specificity. That is, metrics that appear in the literature that are aggregable tend to also be monotonic and those that are monotonic tend to be quantifiable. The practicality of a metric largely depends on what data is available in which these metrics can be evaluated and the complexity of computing the metric (either automatically or manually) or by the complexity of transforming the data into a different representation in which this computation is easier. It should also be noted that these are not always binary characteristics but whether or not they are fulfilled can depend heavily on the clinical context.

Although simply measuring the performance may have scientific value, in order to be used in clinic, a threshold (i.e. *qualifying performance* or M_Q) should be placed on the metrics to determine if the algorithm is of sufficiently high quality to be used. An intuitive example of this could be taken from different cancer biopsy procedures where a particular minimum size is specified for a tumour to be of clinical interest and the error of the system under evaluation must be consistently less than that size [43]. Often in the literature, a comparative approach is used in which another (likely better known or clinically used) model is used and the qualifying performance is implicit, i.e. the performance required for the new algorithm is to simply outperform the old. In many of these cases, M_Q will be described as a distribution rather than a singular threshold, using statistical analysis (discussed in a subsequent section) to determine if the threshold for acceptability is met.

Calculating these metrics often requires a *reference* or *ground truth*, I, i.e. an ideal output value obtained through some other means, such as a more well-accepted process/modality or a human expert. Additionally, the reference representations must be extracted from the clinical data and, although assumed to be completely infallible, they themselves are subject to error. For an image segmentation task, the ground truth may be manual segmentations of the image performed by experts, who themselves vary in how they perceive or conceptualise the underlying anatomy [46]. Again, this is not a binary consideration as even validation methods that may seem fairly direct in applications such as under-sampled medical image reconstruction may rely on ground truth images that themselves have inherent noise and uncertainty. Alternatively, simulation can be used in order to reduce or altogether eliminate the variability and uncertainty in the reference [19].

There are meaningful limits on how well an algorithm can match the reference. These can arise from *reference uncertainty* (M_{RU}) which are errors or other sources of uncertainty in the data used as the reference, or from *observation uncertainty* (M_{OU}) which is noise, errors, or other sources of uncertainty in the input data itself. For an example of M_{OU}, there may be fundamental limits to how well a segmentation or localisation algorithm could perform based on the data's resolution, i.e. differences in distance errors below the size of a pixel/frame are possibly meaningless. A similar argument could be made for the reference itself, although how the reference is generated will often be associated with additional error or uncertainty. These may also be described as distributions if the observations or reference data is highly variable or comes from different sub-types of patients. The degree to which these contextualising aspects are valuable is highly problem dependent. For example, in a signal or image reconstruction task, M_{OU} would be important to contextualise how noise continuously effects the reconstruction, whereas in a discrete classification or segmentation problem, this effect is often assumed to be negligible. On the other hand, for classification and segmentation problems, the quality of the reference information, M_{RU}, has a much higher effect on the perceived quality of an algorithm, especially when it performs well enough to approach the reference itself.

2.5. Validation technique

Once the validation objective and model are defined, the specific *validation technique*, V_T, and *validation granularity*, G_V, can be chosen. (See Table 5 for an example.) The key aspect of V_T is to maintain a "Training/Testing" divide, ensuring that datasets that are used to construct the model are not simultaneously used to validate it. Maintaining this divide is critical as it ensures that the system is put in a context of prospective application on new, unseen patients or in new contexts such as application in a different clinical centre.

Depending on the clinical context (such as the number of datasets available), a number of validation techniques are possible although they largely fall under two categories: *hold-out* approaches and *cross-validation* approaches. Hold-out approaches tend to be simpler and involve taking a particular fraction of the clinical dataset and putting it aside to be used solely for validation purposes. The benefit of this approach is its simplicity: only a single model needs to be constructed and it is always clear where the "Training/Testing" divide lies. In cross-validation approaches, such as leave-one-out cross-validation or k-fold cross-validation, the validation process is repeated multiple times: each time constructing a "Training/Testing" divide, setting F_P based on the data in the current 'Training' set and computing the validation metrics across the data in the current 'Testing' set. The benefit of this approach is that generally more data can be used to determine F_P at a given time, which is important when data is limited. In addition, cross-validation approaches produce more accurate measures of the algorithm's performance on a particular dataset than the hold-out technique given very mild assumptions about the data characteristics and algorithms involved [8]. Thus, cross-validation approaches are more common in MIC than other fields where datasets tend to be larger. However, in practice it is sometimes difficult to ensure that the "Training/Testing" divide is always respected as it is no longer constant over the course of the experiment. Because cross-validation leads to repeating the process of setting F_P multiple times, there is a trade-off between the time taken to perform

Table 5
Example validation technique from Ref. [69].

Validation Technique	
V_T	Direct (i.e. no learned component)
G_V	Patient-level

the validation and the quality of the validation results. For the comparison of two or more algorithms (such as when M_Q is defined by an existing comparative approach) cross-validation also allows for paired statistical testing by matching the 'Training' and 'Testing' splits for all of the algorithms involved, limiting variability that may arise from different "Training/Testing" divides [5].

Nested cross-validation (NCV) is an example of the general versatility of cross-validation techniques. In NCV, the 'Testing'/'Training' divide is creating through the outer cross-validation loop, then the 'Training' dataset is split again in an inner cross-validation loop, providing two sets of data, one of which could be used to train some elements of F_P whereas the other can be used to estimate the performance to train others (generally, the hyperparameters) or to select an optimal model to then evaluate on the actual 'Testing' data. Multiple studies have found NCV to be beneficial in controlling bias for models where there are data normalisation or other hyperparameters that are particularly sensitive to the training data [75,78].

Both validation techniques have issues in estimating the expected performance of the model constructed using a new dataset with the same distribution, the cross-validation technique underestimating its variability due to the high correlation between the 'Training' sets being used and the hold-out technique not estimating a variance for this at all [5]. The impact that this has on algorithms in MIC would seem to indicate that the validation being performed is only representative of that particular model on that particular dataset rather than a family of similar models trained on similar datasets [11], emphasising the need for researchers to not only release the code to train and apply a model, but also the specific instantiation of the model being validated. This naturally raises patient privacy and anonymity concerns if information about a particular patient can be extracted from the resultant published model.

If the MIC algorithm does not involve any trained parameters (i.e. F_P is empty), a common validation technique is to use the entire dataset as 'Testing' data, i.e. *direct* validation. This is more common with non-machine-learning based approaches and is still common in the literature, especially at higher L_A's. It also implies that any configurable weights and hyperparameters of the model are considered part of F_D and should be set without access to any of the data. They should therefore take on pre-defined default values, values that can be reasonably predicted from an understanding of the clinical context and/or data, or values previously reported in the literature on a different dataset. Intuitively, this technique is more common at higher L_A's when the trainable parameters of the algorithm are more likely to be already set and validated in earlier experiments. If direct validation is used, it should be clearly stated, allowing the reader to confirm that no training takes place.

At this stage, it is also important to revisit the idea of the data granularity, finding the appropriate level for the validation as a whole [3]. The *validation granularity, G_V,* is highly important and should be explicitly stated and justified. Although the granularity is often specified (implicitly or explicitly) for the application of the algorithm and for metric computation, it is less frequently noted for the validation as a whole. The G_V for dataset splitting is particularly important as it leads to large differences in quantitative results. Specifically, splitting the data at too fine G_V can lead to data leakage as necessarily correlated datapoints find themselves on separate sides of the 'Training'/'Testing' divide. Depending on the L_A, the different biases caused by different data splitting G_V could be important. For example, images that are taken from the same centre have some degree of correlation, having been acquired by the same device with the same parameters and possibly by the same technician, with patients having more similar geographic, socio-economic, and potentially ethnic/racial distributions which validation at a higher L_A may want to be more robust to and will split the data with a very coarse G_V. At a lower L_A, showing technical feasibility is more important than controlling potentially subtle biases like these.

However, even at alower L_A, the data splitting G_V should not be finer that the patient-level (that is, all the data arising from a single patient must lie on a single side of the 'Training'/'Testing' split) as violating this is well-known to grossly overestimate measures of accuracy [62]. Unfortunately, this level of attention is not always given and, in some fields, only half of the machine learning methods proposed meet this minimal threshold [55].

2.6. Dataset splitting

Once the validation technique has been determined, the clinical data can be processed for the experiment in terms of splitting (which datasets are used for training) and augmentation/sampling (how those datasets are used) as shown in Fig. 5. Splitting in particular consists of the partitioning of the dataset into (potentially multiple instances of) 'Training' and 'Testing' sets according to specified validation granularity. At this stage, the amount of data used for training, N_{train} and for testing N_{test} can be determined. This should be stated at the training granularity and validation granularity at minimum in order to interpret the results both in terms of the data consumption required by the framework and the certainty of the distributions constructed in the validation process. The purpose of applying the model to 'Training' data is to determine or to refine the model parameters, which is often done in an iterative manner. The application of the model to 'Testing' data is to create the output to compare against the reference data and is more representative of how the algorithm is to be used in clinic (Fig. 5).

One key element of the creation of these datasets is ensuring that there is no *data leakage*, that is, information from the 'Testing' dataset that is available to the model as it is determining the model parameters and thus could influence these parameters and bias the results. The most obvious form of data leakage is actively using the same dataset in both 'Training' and 'Testing' datasets, which has occurred in the literature, especially those that use hyperparameter optimisation. This type of data leakage is simple to avoid by assuring that any data splitting occurs before the hyperparameters are determined, even if they are constant.

However, avoiding data leakage in general can be more difficult and nuanced. Of particular importance in detecting and preventing data leakage is considering the granularity described in Section 2.5. A more subtle form of data leakage frequently occurs when only part of a medical image (i.e. a slice or patch) is used by the model at a given time, thus decoupling the notion of a datapoint from a patient. Thus, data from the same patient could be split into both the 'Training' and 'Testing' datasets by having some slices/patches in both sets. This indirectly provides some information about the 'Testing' set to the algorithm during training. The more adjacent or similar these slices/patches are, the more data leakage occurs as the shared information between these datapoints is higher. This can also lead to extremely overestimated performance measurements [62,82]. Depending on the level of assessment, an even greater granularity may be necessary for evaluation, splitting the dataset by centre rather than patient for example to analyse robustness across independent centres [4].

Lastly, if the underlying algorithm or validation technique has non-deterministic components, it may be beneficial to repeat the validation process altogether a certain number of times, V_R. By repeating the validation procedure, the validation results are more reproducible as their variability arising from the underlying non-determinism can be estimated. This helps clarify to those reproducing the algorithm if the result they get is within a reasonable range in a similar clinical context. In the case of machine learning methods, it is especially important for

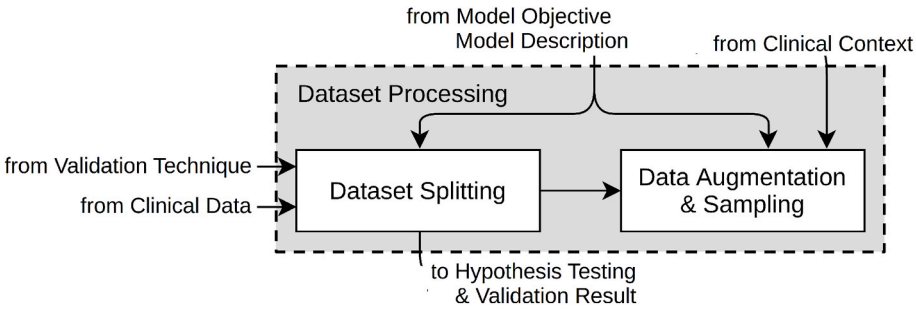


Fig. 5. Dataset processing containing Dataset Splitting, which dictates which sets are used for training or for evaluation of the model, and Data Augmentation & Sampling, which directs how the training data is accessed and extended.

Table 6
Example dataset splitting from Ref. [69].

Dataset Splitting	
N_{test}	10 patient images
N_{train}	0 - no learned component
V_R	No repetitions

these repetitions to be independent and not to share model parameters and thus cause data leakage.

(Examples of N_{train}, N_{test}, and are given in Table 6.)

2.7. Data Augmentation & Sampling approach

For more complex models, the amount of 'Training' data is increased through a *data augmentation approach*, F_A, in which a computational method is applied to a datapoint in order to change it in a non-insignificant manner. For computer vision, this often involves translating, rotating or flipping the image and in medical images, more complex deformations can also be applied depending on the clinical context. These transformations may also be chosen to reflect possibly missing properties of the model itself. For example, if the model is not inherently invariant/equivariant to rotation and translation but should be, introducing rotation and translation into the data augmentation approach may allow it to learn that invariance in a robust manner rather than having it hard-coded.

A crucial aspect of data augmentation is ensuring that the datapoint stays consistent under the applied transformation. That is, if a transformation is applied to one aspect of the datapoint then related aspects of that same datapoint should change too. For example, if our goal is to estimate the volume of a particular organ of interest, and part of your F_A involves spatially scaling the image, the same scale should apply to the volume. For some problems, such as diagnosis problems, certain perturbations of the data intuitively do not require any further changes to render them consistent (e.g. if you are diagnosing a disease, shifting the entire image by a few pixels does not change the disease state, but deformation might). Another key element to consider for F_A is what forms of data augmentation are clinically feasible. For example, left-right flipping of brain images for segmentation problems is relatively common, with the differences between the left and right hemisphere being insignificant compared to other sources of error. This however is not the case with diagnosing particular neurological disorders in which there is a defined left-right difference. In a more extreme example, flipping should never be used in abdominal image processing due to its inherent lack of symmetry, except for the detection of *situs inversus* in which such flipped images are not abherrations but completely warranted given the context. The scale of the data augmentation should not go beyond the point where the augmented data is no longer representative of a reasonable clinical case.

It is important to note that any form of data augmentation has the potential to introduce some level of bias, A_B. For example, rotating an image may sound bias-free, but can introduce small interpolation errors when the image is resampled which may effect both the algorithm and the reference. Although data augmentation is used principally on the 'Training' data, there are also instances in which it may need to be applied to the 'Testing' data as well in order to validate that the in/equi-variance properties do hold or to get a measure of the M_{OU}. In these cases, a critical examination of potential biases is especially important.

Increasingly, data augmentation via a separate generative machine learning process, such as generative adversarial networks (GANs) [13,20] or variational autoencoders [56], is being used in order to capture more complex invariances that are difficult to capture or to model using traditional descriptions. These data augmentation techniques can be used in tandem with some caveats, such as the possibility for compounding error and generating statistically unrepresentative datapoints [72]. In these cases, it is often not possible to specify an exact in/equi-variance is being targeted, implying that a validation of F_A itself may be necessary to ensure it fits the intended clinical context.

Complex data aumgnetation methods may also be a source of data leakage if not carefully considered. That is an element from the 'Testing' set could have been in the data augmentation method's 'Training' set. This means that the data augmentation could generate datapoints that are very similar to the 'Testing' dataset only by virtue of having prior access to them. This again is easy to avoid by ensuring that data augmentation is considered part of the method as a whole (i.e. data splitting is performed prior to training the data augmentation method) or that the data augmentation method is developed on an entirely independent database, even if it was created by an independent research group.

Another element that can have a strong effect on the validation of the model is how the 'Training' data is sampled during the process of determining the model parameters. This *sampling approach*, F_S, is often fairly simple including a (repeated) randomisation of the training data order ensuring that all the 'Training' data is used, but that spurious correlations arising from the order in which the data is presented are minimised. Some more complex adaptive approaches have also been developed, which allow for the algorithm to increasingly focus on what it empirically finds to be difficult cases [16,53].

(An example of these variables in given in Table 7, noting that data augmentation did not appear to be used in the cited paper, thus potential methods are given.)

Table 7
Example data augmentation and sampling approach using a registration study for image-guided needle biopsy for prostate cancer [69].

Data Augmentation & Sampling	
A_{IE}	None specified - possible invariance to translation, rotation, scale, noise, etc …
A_B	None specified - possible interpolation errors
F_A	None specified - possible application of rigid transformations, scaling, noise addition, etc …
F_S	Uniform

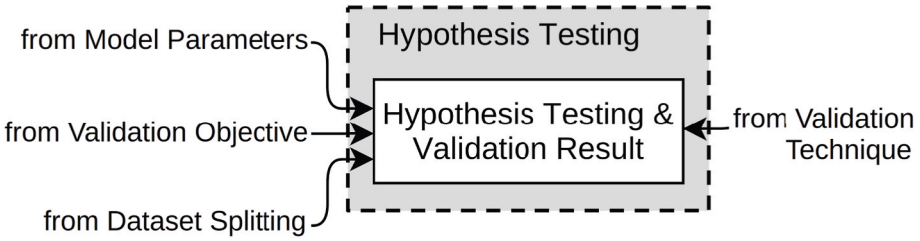


Fig. 6. Hypothesis Testing & Validation Result describes how the performance of the system is quantified, reported, and then compared to a pre-defined level or hypothesis.

2.8. Hypothesis verification & the validation result

In terms of the validation itself, the crucial terminal aspect is the reporting and verification of the results as shown in Fig. 6.

In general, *hypothesis verification* is well-understood and papers already report.

- M_O, the *observed distribution* of results,
- F_H, the method for comparing M_O to M_Q, often using a statistical framework.

(An example of an instantiation of these variables is given in Table 8.)

Statistical methods principally work by estimating the probability of generating the observed distribution under some null-hypothesis, that only just barely misses the qualifying performance. There is a large degree of literature about different statistical tests to perform, with the most common largely being based on *t*-test, paired *t*-tests (for when the qualifying performance varies across patients, such as when a comparative method is used) or *F*-tests for factorial experiments. These tests make certain assumptions regarding properties of both M_O and M_Q, such as normality and homoscedasticity, although the degree to which these assumptions can be violated while still rendering the test valid is still debated. Non-parametric tests that use permutation testing and rank statistics have weaker distributional assumptions but they are less commonly used, less flexible, and are generally more conservative. Ultimately, the result of the statistical test is the validation result, i.e. whether or not M_O falls in an acceptable range. Another element of the statistical framework to report is the correction method (if any) for multiple comparisons when M_V has multiple components.

Statistical correction may also be necessary depending on the number and type of validation metrics. Although simple methods such as Holm-Bonferroni correction work for a small number of metrics, more complex corrections may be needed for certain problems in which there are a large number of correlated metrics (for example, voxel-based analysis [12,54]). In some cases, authors may choose to forego statistical testing altogether. This is becoming increasingly common in machine learning methods in which the act of training a model itself is time consuming and thus it may not be feasible to collect the number of measurements of M_O necessary to make statistical conclusions. As one should critique papers that use statistical methods based on the appropriateness of the method used, the same should be applied to those that do not use any statistical technique; some additional justification should be given as to if the qualifying performance is met and under what conditions.

Table 8
Example hypothesis verification using a registration study for image-guided needle biopsy for prostate cancer [69]. Quotations are taken directly from the text. Non-quotations are either implicit in the text or are generated for the sake of providing a reasonable example.

Hypothesis Testing & Validation Result	
M_O	"TRE values of 1.97 mm, 1.58 mm and 1.74 mm respectively", "the proposed method generated a favorable DSC value of 92.9 ± 2.6% for the mid-gland, 83.0 ± 5.6% for the apex, and 80.1 ± 4.7% for the base.", "The mean registration time of our method per patient was 90 ± 5s in addition to 30 ± 5s for initialisation."
F_H	Not statistically defined but frequency-based - "Fig. 3(b) shows that 80% of the TRE values for WG and 76% for PZ are below the desired values."

2.9. Resultant model

The other terminal aspect is the model/algorithm to be used itself as shown in Fig. 7. For most traditional algorithms, the model to be used (i.e. the *resultant model*) is often the exact same as the model being validated, i.e. *direct use*. This is also generally more common at medium-high and high L_A's where the model is more likely to be considered fixed entirely by F_D with no F_P. However, machine learning models require a more nuanced approach and thus the paper should specify its method of creation, F_R. (An example of an instantiation of this variable is given in Table 9.)

Even for simple machine learning methods, the final model intended for use may not be any of the models constructed during the validation process, but generated from them or from the entirety of the data. This is more common for validation techniques that generate multiple models, such as cross-validation. In these cases, model ensembling is often used, which creates a new model by aggregating the predictions of the models constructed during the experiment. Certain simple forms of ensembling, such as averaging the output of several equivalent models (i.e. bagging), are known to have equivalent if not better accuracy than the models in the ensemble at the expense of requiring more time and memory.

For simpler models with relatively few trainable parameters or ones in which the trainable parameters have a known, modelable effect, the final model can be constructed simply retraining the model using all available data can be done. This has the benefit of using the same model in application as in validation, thus keeping aspects such as time and memory closer to their measured values, while also making use of all of the available data. Interestingly, retraining does not guarantee equal or improved performance, especially more complex deep learning models in which *double descent* may apply [52]. For certain models, increasing the amount of training data without changing its structure can actually reduce its performance.

The construction of the final model is not without some nuance as it must still be compatible with the validation result. For example, a common implicit form of resultant model construction is the selection of the best performing model based on the validation experiment. However, it should be noted that this selection itself is a source of a slight bias as the

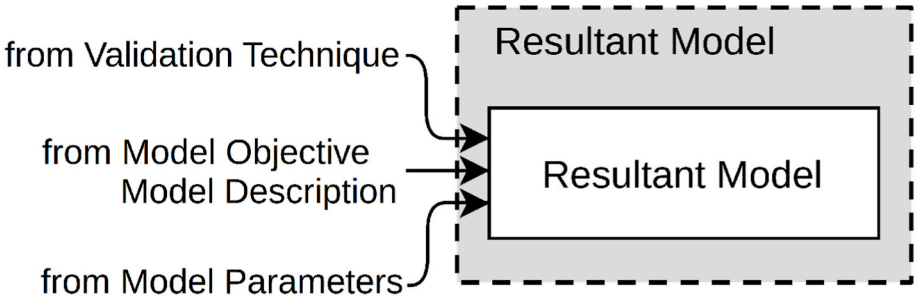


Fig. 7. Resultant Model describes how the model to be used is or can be constructed from those investigated during the process of validation.

Table 9
Example resultant model construction using a registration study for image-guided needle biopsy for prostate cancer [69].

Resultant Model	
F_R	Direct use

reason why the model has the best performance could be due to random variations. This is exacerbated as more models are included in the analysis, especially if they have similar performance and thus their relative differences are lessened in comparison to random variation [6, 38].

3. Literature review examples

One of the primary uses of the proposed framework is to structure elements of literature reviews in order to make them more sensitive to how the algorithms under investigation are validated and interpreted in comparison to each other. Our recent literature [55] incorporated elements of an early version of this framework, using them to find that over half of machine learning methods used in the context of deep brain stimulation involved potential data leakage between 'Training' and 'Testing' datasets. However, the framework was not the focus of said article and its instantiation wasn't shown in great detail. The following small literature reviews are therefore not designed to be exhaustive in terms of the literature, but to give a more complete example of the framework in use. For both, we have attempted to get a variety of techniques spanning the previous decade in order to illustrate the general usability of this framework to different paradigms in their respective fields.

Elements of the clinical context and data are used to limit the scope of the literature being reviewed, ensuring the methods being compared are attempting a similar problem and could in theory be used interchangeably. The remaining categories form the basis for how the different methods in the literature differ from each other and are to be critically appraised. Specifying the L_A in this case is crucial to best control the scope of the techniques under investigation while not overlying constraining them to look at a particular validation objective. This also allows for research systems, which are generally validated at lower L_A's, to be easily distinguished from systems that are in current clinical use where higher L_A's are more common.

We have structured the literature reviews in a tabular format following the notation established in Section 2. This allows for validation information to be reported in a systematic manner. Depending on the subject, the information regarding the clinical context, clinical data, and model structure used could be greatly expanded cognisant of the particular literature being investigated. For example, if the literature review concerns MR imaging, it could include the scan parameters as a dedicated column. For investigating the use of a particular class of models, F_O, F_D, and F_P may need to be expanded to identify particular architectures. Despite this, our proposed framework still necessitates that these added columns be clearly delineated into the different categories, particularly for the model structure in which there is a strong distinction between parts of the model that are fixed prior to the experiment (i.e. F_O & F_D) and parts that are affected by the particular data used (i.e. F_P). This is because maintaining this separation is particularly important to maintaining a 'Training'/'Testing' divide.

3.1. Segmentation of the subthalamic nuclei in MRI

Deep brain stimulation (DBS) of the subthalamic nucleus (STN) is an increasingly common treatment for drug-resistant Parkinson's disease. During this procedure, electrodes are placed within the STN, which requires it to be segmented during the pre-operative planning stage in order to develop appropriate electrode trajectories. As the subcortical anatomy in this region is quite complicated with several structures of interest, automatic segmentation methods are sometimes used. However, due to the small size of the STN and its low contrast with surrounding tissue on T1-weighted imaging, these methods need to be highly accurate and robust to noise.

Given the clinical context outlined in Table 10, which was designed to specify comparable clinical scenarios and thus methods that could potentially be used interchangeably, a Google Scholar search was performed (search string: *"subthalamic nucleus" "MRI segmentation" "deep brain stimulation"*, date range: 2015-present, accessed November 15, 2022) and the first seven methods selected (chosen to ensure the tabular results fit on a single page) which all occurred in the first 50 results and more than half occurring within the first 10 results. Some methods were

Table 10
Clinical context for the STN segmentation literature review.

Clinical Context	
C	MRI-based pre-operative planning for deep brain stimulation of the STN
L_A	Medium-low level - accuracy on clinical dataMedium level - margin of error in clinic
L_I	Fully automatic
T_H	Visualisethe location and extent of the STN in a pre-operative MR image
T_I	Image selection only (i.e. no initialisation/correction)
T_A	Segmentation of the left and right STN
Clinical Data	
P_D	Singular clinical quality (i.e. 1.5T or 3T) T1w and/or T2w MR scans (res. approx. 1 mm iso.)

very close to meeting the criteria, such as Varga et al. [77] which had a small manual initialisation step but was otherwise automatic. Another interesting paper [57] implemented and compared several techniques, benchmarking their performance as well as providing metrics of inter-rater variability for manual STN segmentation.

The results of the literature review are given in Tables 11 and 12, with the first table reporting the context and methodology with the second reporting the more validation-centric criteria. From a methodological point of view, one particular group [32–34,63] was dominant, using a 7T atlas (with deformable registration and some post-processing) as their method. However, their papers varied in terms of validation, notably moving from a medium-low L_A in Refs. [32–34] to a medium L_A in Ref. [63] as the algorithms were no longer validated against manual segmentation but against electrophysiological recordings taken within a DBS intervention. Reinacher et al. [59] used the same medium L_A with a very similar validation objective, receiving extremely comparable results to Ref. [63] despite using a completely different algorithm for STN segmentation. Li et al. [36] attempted to bridge the two L_A's by not only comparing against electrophysiology, but also having a manual segmentation which acts as a reference as well as a M_Q for the medium L_A. However, the limited data used as well as the strong assumption that brain-shift is negligible makes it more difficult to draw conclusions about the relationship between segmentation overlap and agreement with electrophysiology. The remaining papers [4,47] used similar metrics which would make them appear similar at a surface level, but [47] uses a reference defined by an atlas. From a data point of view, the two methods also differ greatly in terms of granularity, with one treating patches as the fundamental unit of data for network training and the other using the entire volume at once) (see .

One interesting point for the papers [4,33,36] that used manual segmentation as a reference is that none provided a measurement of M_{RU}. For this particular clinical context and problem, M_{RU} for manual segmentation has since been estimated to be on the order of a 63% Dice co-efficient [57] which adds credence to methods with a similar performance [4,33] while questioning those with much higher reported accuracy [36]. For the methods that used deformable registration as their reference [34,47], again, M_{RU} was not provided. This can be problematic as atlas registrations are generally less variable but higher error than manual segmentation, meaning that an increase in Dice performance could be due to the erroneous consistencies of atlas-based segmentation. The fact that all of these papers lack a reference uncertainty and most lack an observation uncertainty shows how unfortunately uncommon these contextualising pieces of information are in this particular subfield.

In terms of statistical testing, only half the papers [33,59,63] explicitly provided information on F_H; one paper [4] gave enough information for a two-tailed t-test to be performed, and two [34,47] did not give enough information for any statistical test to be performed. Again, this may call into question to what degree the methods presented achieve their M_Q.

The use of our framework makes it easier to compare and contrast papers that are extremely different in their methodology, noting that the algorithms used included deformable registration [33], level-set segmentation [36], shape/transform regression [33,34,63], and completely different types of convolutional neural networks [4,47].

Table 11
Papers used in STN segmentation literature review - Part 1.

Paper	*Clinical Data*				
	P_P	P_I	S_{ID}	S_{OD}	A
Baxter and Jannin [4]	Parkinson's disease patients (6 M+4F)	Intensity normalisation (divide by 95 percentile) and smart cropping	T1/T2 pair	Binary mask	Images are perfectly coregistered
Kim et al. [33]	Parkinson's disease patients	Registered to 7T atlas	T1/T2 pair	Mesh	7T registration does not fail
Kim et al. [34]	Mixed Parkinson's, tremor patients and healthy subjects (61 M + 19F)	*N/S*	T2 image	Mesh	Accurate 7T image registration
Li et al. [36]	Parkinson's disease patients (6 M+4F)	*N/S*	T2 image	Mesh	STN in central image location
Milletari et al. [47]	*N/S*	*N/S*	SWI image	Binary Mask	*N/S*
Reinacher et al. [59]	Parkinson's disease and dystonia patients	*N/S*	T1, T2, images	*N/S*	Brain shift is negligible, electrical recordings well registered
Shamir et al. [63] (method: [32])	Parkinson's disease patients (8 M+8F)	*N/S*	T1w image	Mesh	Brain shift is negligible, electrical recordings well registered

Paper	*Model Structure*				*Resultant Model*
	G_F	F_O	F_D	F_P	F_R
Baxter and Jannin [4]	Image pair	Minimize voxel-wise BCE	U-Net architecture, number of layers, widths, etc …	Network weights	Voxel-wise voting
Kim et al. [33]	Image	Align points and maximise information gain	Procrustes, kernel PCA, regression forests to find transform	regression tree weights, kernel PCA transform	Direct use
Kim et al. [34]	Image	Maximise shape model fit	Regression-based shape prediction (active contours + active shapes) + ensembling	Active contour + active shape params, ensemble weights, etc …	Direct-use
Li et al. [36]	Image	Level set functional	Initial thresholding/cropping, level-set opt., spline mesh fitting	*N/S*	Direct use
Milletari et al. [47]	Patch (CNN) & image (post-processing)	Custom Dice-based loss (seg.), *N/S* (voting)	Multiple comparative architectures	Network weights	Selection of best-performing architecture
Reinacher et al. [59]	Image	*N/S*	Surgiplan software	*N/S*	Direct use
Shamir et al. [63] (method: [32])	Image	Maximise shape model fit	Regression-based shape prediction (active contours + active shapes) + ensembling	Active contour + active shape params, ensemble weights, etc …	Direct-use

N/S meaning *not specified* i.e. said information is not present in the paper, *N/A* meaning *not applicable* i.e. when said information is not sensible in the validation context.

Table 12
Papers used in STN segmentation literature review - Part 2.

Paper	*Validation Objective*						*Validation Technique*		*Data Splitting*		
	V	M_V	I	M_Q	M_{OU}	M_{RU}	G_V	V_T	N_{test}	N_{train}	V_R
Baxter and Jannin [4]	· overlap	Dice coeff.	Manual segmentation	Within M_0 (implicit)	1 mm shift/ dilate (50.5 ± 2.8% to 55.0 ± 5.5%)	*N/S*	Patient-level	LOOCV	10	10	5
Kim et al. [33]	· distance · orientation · overlap	Dice, rMS orientation error, CoM distance	Manual segmentation	Compare to 7T atlas	*N/S*	*N/S*	Patient-level	'Testing' set	36	10	1
Kim et al. [34]	· distance · overlap	Dice, CoM dist., MSD	Registered 7T manual segmentation	Compare to registration methods	*N/S*	*N/S*	Patient-level	LOOCV+ 'Testing' set	92	46 + 34	1
Li et al. [36]	· overlap · distance	Dice, dist. between *positions* (undefined)	Manual segmentation, registered electric signals	overlap: *N/S* distance: equivalent to expert	*N/S*	*N/S*	Patient-level	'Testing' set only	*N/A*	10	1
Milletari et al. [47]	· distance · overlap · robustness	Dice, mean contour dist., failure rate	Atlas registration	Compare to traditional CNN	*N/S*	*N/S*	*N/S*	*N/S*	135k to 13.5 M	10	8 or 9
Reinacher et al. [59]	·distance	Distance along trajectory	Registered inter-op electrical recordings	Negligible difference from I	*N/S*	*N/S*	Patient-level	'Testing' set only	*N/A*	30	1
Shamir et al. [63]	·extent	electrode trajectory length in STN	Registered inter-op electrical recordings	Equivalent to reference	*N/S*	*N/S*	Patient-level	'Testing' set only *(pre-trained)*	*N/A*	16	1

Paper	*Data Processing*		*Hypothesis Verification*	
	F_A	F_S	M_O	F_H
Baxter and Jannin [4]	Rotation and translation	Uniform	58.2 ± 12.1%	*N/S*
Kim et al. [33]	either none or *N/S*	Uniform	1.2 ± 0.5 mm (left) 1.2 ± 0.5 mm (right) 13.1 ± 6.4° (left) 17.6 ± 12.9° (right) 61 ± 12% (left) 56 ± 15 (right)	*N/S*
Kim et al. [34]	either none or *N/S*	Uniform	1.28 ± 0.65mm62 ± 13%	One-way ANOVA with multiple comp. correction, Tukey post-test (sign. 0.1%)
Li et al. [36]	either none or *N/S*	*N/A*	88 ± 4% (left) 86 ± 5% (right) 1.1 ± 0.6 mm (left) 1.1 ± 0.7 mm (right)	overlap: *N/S* distance: *N/S*
Milletari et al. [47]	either none or *N/S*	Uniform	*multiple results due to multiple architectures*	*N/S*
Reinacher et al. [59]	either none or *N/S*	*N/A*	0.6 − 1.9 mm (entry) 1.35 − 3.0 mm (exit)	Two-tailed *t*-tests
Shamir et al. [63]	either none or *N/S*	*N/A*	5.8 ± 0.9 mm (versus 6.2 ± 0.7 for electric signals)	Paired *t*-tests (sign. 5%)

N/S meaning *not specified* i.e. said information is not present in the paper, *N/A* meaning *not applicable* i.e. when said information is not sensible in the validation context.

3.2. Point localisation for transcranial magnetic stimulation

Repetitive transcranial magnetic stimulation (TMS) is a therapeutic technique for various neurological disorders due to its ability to affect and disrupt pathological cortical behaviour. As early as 2001, studies [22] have found that the traditional methods for guiding these procedures in clinic are inaccurate and have advocated for personalised image guidance based on MRI. This involves several components, the first being the registration of the TMS probe into the co-ordinate space of the patient image, allowing for *visually guided TMS* which was shown to improve the consistency for finding cortical points in the motor region with the help of electromyography feedback [18,28]. However, there was still a question of whether or not this could be further improved by automatically finding the target stimulation sites in pre-operative images rather than through this feedback mechanism. With the growing popularity of machine learning algorithms, this latter step could be automated, with a computer determining stimulation sites based immediately on the patient image. Functional MRI images soon became the standard for identifying TMS cortical target points [67], although it was less useable in clinic due to the additional scanning requirements. This review looks at the papers that propose automatic cortical point localisation for TMS from only T1-weighted MRI representative of clinical use.

As with the previous subsection, the specific clinical context of the literature review is presented in Table 13 using the symbols suggested in the previous section. For the human task, we have specified that the task to the pre-/peri-operative determination of the location of potential stimulation sites, rather than the intra-/post-operative task of coil positioning or determining what sites is stimulated or the selection of a

Table 13
Clinical context for the TMS point localisation literature review.

Clinical Context	
C	Pre-/peri-operative planning for therapeutic repetitive transcranial magnetic stimulation, specifically for neurological disorders with stimulation sites in the frontal cortex
L_A	Medium-low level - accuracy on clinical data Medium level - immediate effect on symptoms
L_I	Fully automatic
T_H	Determining the position of potential stimulation sites given an unannotated T1-weighted MR image
T_I	Image selection only (i.e. no initialisation/correction)
T_A	Estimation of a fixed-length vector of points coordinates in image-coordinate space
Clinical Data	
P_D	Clinical (i.e. 1.5T or 3T) T1w MRI (res. 1 mm iso.)

particular site based on the patient's symptoms, both of which are also investigated in the literature but address fundamentally different problems.

By specifying the level of interaction as being fully automatic, we limited the type of systems under investigation to ones that are interchangeable from a clinical workflow perspective. Due to the specificity of process, only five papers filling these objectives were identified in the first two hundred results from a Google Scholar search (search string: *"transcranial magnetic stimulation" MRI navigation automatic*, date range: 2001-present, accessed November 15, 2022). An additional paper

Table 14
Papers used in TMS point localisation literature review - Part 1.

Paper	*Clinical Data*				
	P_P	P_I	S_{ID}	S_{OD}	A
Baxter et al. [1]	Chronic pain and depression patients	Intensity normalisation (divide by 95 percentile) and resizing	RAS images	Point vector	*N/S*
Reijonen et al. [58]	Patients including those with depression and schizophrenia	*N/S*	Image	Painted mesh	*N/S*
Rusjan et al. [60]	Healthy subjects (10 M+5F)	*N/S*	Images	Point Vector	Negligible tracking error
Sparing et al. [67]	Healthy subjects (6 M+4F)	*N/S*	Images	Heat Map	Probe positioning is done with zero error
Zosso et al. [85]	Images from patients who have undergone TMS	*N/S*	Images	Point vector	*N/S*

Paper	*Model Structure*				*Resultant Model*
	G_F	F_O	F_D	F_P	F_R
Baxter et al. [1]	Image	Minimize error on a training dataset	· CNN with a fixed structure· Layer resolutions, loss function parameters, etc …	Network weights	Bagging (ensembling by prediction averaging)
Reijonen et al. [58]	Image	*N/A*	FreeSurfer cortical reconstruction and Brainnetome atlas fitting	*N/S*	Directed use
Rusjan et al. [60]	Image	*N/S*	SPM2 registration to Talairach atlas	*N/S*	Direct use
Sparing et al. [67]	Image	Minimize registration cost between patient and ICBM152 atlas	Atlas heat-map determined by population fMRI	*N/S*	Direct use
Zosso et al. [85]	Image	Minimize mutual information between a registered image and an atlas image	· Rigid registration · B-spline non-rigid registration	*N/S*	Direct use

N/S meaning *not specified* i.e. said information is not present in the paper, *N/A* meaning *not applicable* i.e. when said information is not sensible in the validation context.

Table 15
Papers used in TMS point localisation literature review - Part 2.

Paper	*Validation Objective*						*Validation Technique*		*Dataset Splitting*		
	V	M_V	I	M_Q	M_{OU}	M_{RU}	G_V	V_T	N_{test}	N_{train}	V_R
Baxter et al. [1]	·distance ·consistency	· mean dist. ·st.dev.	Manual (3 experts)	Compare to reg. 6.38 ± 3.30 mm to 13.32 ± 3.33 mm	*N/S*	5.65 ± 3.95 mm to 8.84 ± 5.45 mm	Patient-level	LOOCV	26	26	5
Reijonen et al. [58]	·distance	shortest distance to boundary	Expert selection, TMS mapping, and [50] (DLPFC)	*N/S*	*N/S*	*N/S* for expert & mapping, otherwise in [50]	Patient-level	'Testing' set only	*N/A*	22	1
Rusjan et al. [60]	·distance ·consistency	·mean dist ·tracking stdev. in MNI space	Experimentally determined point	Compare to blind methods	*N/S*	*N/S*	Patient-level	LOOCV	15	15	1
Sparing et al. [67]	·immediate physiolog- ical effect	evoked motor potential amplitude	*N/A*	Compare to blind/ manual/ registration methods	*N/S*	*N/A*	Patient-level	'Testing' set only	*N/A*	10	1
Zosso et al. [85]	· distance ·consistency	· inter-means ·st.dev.	Mean of comparative registration locations	*N/S*	*N/S*	7.96 to 9.22 mm	Patient-level	'Testing' set only	*N/A*	10	1

Paper	*Data Processing*		*Hypothesis Verification*	
	F_A	F_S	M_O	F_H
Baxter et al. [1]	Translation (std. 10 mm) Rotation (std. 10°)	Uniform	distance: 5.54 ± 3.25 to 9.08 ± 8.29 mm consistency: 1.33 ± 0.93 to 3.24 ± 5.60 mm	Wilcoxon test for mean error, sign. 5% Bonferroni correction
Reijonen et al. [58]	none	*N/A*	distance: 5 ± 2 to 23 ± 4 mm	none
Rusjan et al. [60]	none	*N/A*	distance: 12.6 ± 6.7 mm consistency: *only in figure*	Wilcoxon signed-rank test
Sparing et al. [67]	none	*N/A*	relative ampl: 72 ± 7%	One-way repeated measures ANOVA, sign. 0.1%
Zosso et al. [85]	none	*N/A*	distance: 5 to 7 mmconsistency: 7.8 to 11 mm	none

N/S meaning *not specified* i.e. said information is not present in the paper, *N/A* meaning *not applicable* i.e. when said information is not sensible in the validation context - (st.dev) standard deviation of predicted co-ordinates.

was found that was close to meeting the requirements, but had a higher L_A involving a medium-term clinical trial [28].

The results of the literature review are given in Tables 14 and 15. In terms of the methods used, two of the three papers made use of deformable registration and an atlas in order to determine the location of the potential TMS stimulation points. Table 14 contains information about the general context of the experiments and the specific algorithms implemented (and, for a long literature review or a more method-focused literature review, it should be further extended and sub-divided based on more specific elements of the methodology). Interestingly, every paper used atlas registration either as the proposed method [58,60,67,85] or as the comparative method [1]. In terms of data granularity, each method used a single T1-weighted image per patient and the algorithms processed the entire image at once, leading to an equivalence between image-level and patient-level granularity used throughout each method. Interestingly, none of the methods specified the parameters used for registration and how they were determined, which may lead to a lack of reproducibility and potential bias in the results.

As with the previous literature review, the second table (Table 15) is dedicated to aspects of validation and shows much more heterogeneity, even amongst only five methods. For example, for the three methods with the same V (i.e. distance and consistency), M_V is defined in completely different ways due to one having a manual reference [4], one having a reference derived from TMS mapping [60], and one deriving it from more well-known registration method [85]. Although the two of the three methods produce similar numeric accuracies, the numbers themselves are not directly comparable because they only appear to measure the same thing. The method with the numerically worst accuracy uses a more objective and possible higher quality reference, meaning that it may not truly underperform the other two. For the other papers, L_A and V themselves were fundamentally different. For example, Sparing et al. [67] used a more middling L_A, measuring the effect that changing the point localisation method would have on the immediate symptoms by measuring the evoked potential at the desired target anatomy (e.g. the hand) due to stimulating a cortical region.

Again, the papers overall tended to lack M_{RU} and M_{OU} information to contextualise their results. In the case of Baxter et al. [1], M_{RU} was provided, which was used to contextualise the performance of the algorithm in terms of approaching human expert performance. Reijonen et al. [58] used a citation from the literature for M_{RU} [50], although this literature used different validation metrics, making it still difficult to contextualise the results. In the case of Sparing et al. [67], no M_{RU} is needed as a reference-free approach was used for validation. None of the papers defined M_{OU}, as the consistency metric used by Baxter et al. [1] is a function of the random initialisation of the model parameters (something controlled by the training process) rather than randomness or uncertainty in the data itself.

Another interesting point is the number of papers that are missing an aspect of validation that is usually considered critical: a statistical test that gauges if an improvement is large enough as to not be caused by random chance. In both cases [58,85], this was because the papers did not have even an implicit notion of M_Q which is more common for exploratory papers rather than those which validate a technical method. It is not necessarily a negative thing for exploratory papers to be purely observational rather than define a level for clinical use as, at that stage, the technology may be sufficient far from clinical use that immediately comparing the performance with a qualifying level would prevent these explorations from seeing publication, regardless of their technological novelty, which could stifle necessary initial innovation.

3.3. Connections to framework

Aside from their use of Section 2's catalogue of symbols and definitions to both filter and summarise papers, these mini-literature reviews demonstrate the importance of giving a higher degree of attention to validation details. One initial observation is the heterogeneity in which papers are validated as observed in these mini-literature reviews as well as in other validation-critical literature reviews [55,73].

This arises because the validation of medical information processing algorithms does not follow a singular path in practice, but is composed of a series of decisions, techniques, and best-practices that are not always mutually consistent. This inherently poses issues for the MIC community as a whole as it would mean that in addition to different teams having access to different data and expertise in different methodological techniques, the algorithms being produced are not evaluated on a consistent and globally understood manner. Recently, several researchers have openly worried that machine learning in general may be experiencing a *reproducibility crisis* [24,45,68] largely because of this.

4. Improving validation at a community level

As noted by both Jannin et al. [26] and Maier-Hein et al. [41] and is also argued for in this paper, it is clearly necessary to formalise validation procedures and ensure that papers adhere to this formalism,. This could be either by encouraging reviewers to look for all the elements of the formalism, or by necessitating a specific author-filled checklist for submission. This will unarguably improve validation through making the procedures clearer and better reported, it also may draw the attention of people performing the validation to specific sources of systemic bias that can be account for or corrected through relatively small changes to the validation procedure. Other, more community-level approaches could also be taken, using this standardisation as the basis for validation-critical literature reviews, publication of recommended validation approaches, standardised open datasets, comparison papers, and competitive challenges.

4.1. Validation-critical literature reviews

As shown in Section 3, one can easily use this framework to structure information to make validation-critical literature reviews which do not neglect nor blindly accept numeric results reported in papers. Although validation information does not make up all information normally present in a literature review (notably the enumeration and distillation of methods used) it provides another lens through which to critique the literature and the research community. These reviews cannot change the validation quality of past papers but they could provide an opportunity to identify common trends in how algorithms for a particular problem are validated, which may encourage others to structure their own experiments more conscientiously.

Encouraging validation-critical literature reviews is by far the least intrusive way to improve validation quality, however, it is also the weakest as it would still allow for multiple mutually incompatible types of validation for a single technical problem to exist at the same time without clarifying which is best in terms of translating technology into clinical use. In order to address this, it would have to take a normative stance, advocating for a particular validation structure rather than relying on the research community to coalesce around one organically.

4.2. Publication of recommended validation approaches

To proactively standardise validation, a team of expert researchers and clinicians could publish recommendations on what validation configurations are correct, meaningful, and of use to their community and for authors to adhere to these recommendations. This would allow clinicians to specify reasonable limits on the clinical context and data (e.g. what imaging sequences are used for a particular problem, what computational/temporal resources are allotted for a particular computational task) and for domain experts to judge (i.e. recommended validation techniques based on the number of images a centre is likely to have, appropriate validation metrics, etc …). In the approach described by Jannin et al. [26], this would be necessary for almost all levels of

validation standardisation. Some initial forays have been made into this in a general context, notably for selecting appropriate validation metrics in image segmentation and classification problems [40].

By publishing recommended validation approaches that are detailed and precise, the burden (and thus the variability) of developing these procedures is lifted from researchers engaged in developing or validating new technologies. One could imagine the majority of papers could simply cite the guideline being used. In addition, any systemic issues with the guideline could be addressed in a more open manner through meta-analysis papers or through papers that compare these guidelines against a more accurate (and likely more time-consuming) approach.

For papers that depart from these guidelines, the burden is then on the authors to justify differences between their approach and the guideline and to minimize the negative effects of these differences. For example, if the guideline recommends a particular pre-processing method that is unavailable to the researcher for computational reasons, they can substitute it for another, justifying its validity.

Overall, this approach would encourage more *modularity* in terms of submitted papers, no longer requiring extensive detail required to adequately describe their validation approach, but rather a citation. These papers would become a tool for authors and for reviewers for standardisation, facilitating meta-analyses and literature reviews. It would also be more flexible than using a particular published verification phantom or dataset.

However, there are some disadvantages. For many evolving applications, there may not be enough information to determine these validation guidelines, leaving researchers to have to initially develop them on their own. The second issue is that these guidelines must avoid being too restrictive, which could raise the barrier-to-entry for novel techniques, and being too relaxed and not provide a uniform validation approach.

4.3. Standardised open datasets

Possibly the most common way to improve validation in terms of comparability across the published literature is via the creation and distribution of standardised open datasets. Several of these datasets already exist in the medical imaging community as part of larger longitudinal initiatives such as the Parkinson's Progression Markers Initiative (PPMI) [42] or the Alzheimer's Disease NeuroImage (ADNI) database [49]. These datasets are often a result of a large-scale collaboration between multiple centres that have structured methods for ensuring dataset quality and consistency, as well as reporting patient distributions that are representative of the patient population as a whole, bringing these methods closer to clinical use.

One of the issues with standardised open datasets however is that they engender a particular type of bias resulting from their availability to the community and their acceptance as a *gold standard* for comparison. Essentially, an open dataset that is ubiquitously used allows for methods to become tailored specifically to that dataset and not necessarily to clinical use [4]. For example, the Shepp-Logan phantom [64] and its variants [15] have been heavily used in the literature for MRI and CT reconstruction with innumerable articles validated solely using this phantom. But, it is known that this phantom biases methods towards piece-wise constancy [66] with one early method even having zero reconstruction error for the original phantom [74] which obviously cannot happen in a real clinical scenario. For open datasets consisting of real images rather than phantoms, this can be at least partially addressed by periodically adding to the database, creating new indexed versions with more heterogeneous data for more extensive, on-going, and less-biased validation.

The potential contribution of our framework to standardised open datasets would be to again encourage more structured and complete reporting especially in terms of the quality and certainty of the dataset itself. In addition, our framework also suggests ways in which structured open datasets could become more useful, such as the inclusion of data augmentation procedures and guidelines. Although simple data augmentation procedures (e.g. image translation/rotation, etc …) are relatively easy for researchers to construct, more problem-specific or annotation-specific augmentation capabilities could be envisioned. For example, certain problems such as medical image segmentation and registration, have invariances regarding deformable transformations. However, for some anatomies, such as the brain, what deformations are possible is itself constrained (i.e. by brain symmetry) and the plane of symmetry could be an annotation provided by the group constructing the dataset which would be more standardised and efficient than having the groups that use the dataset develop their own annotations.

4.4. Comparison papers

Comparison papers are similar to standardised open datasets in that a singular dataset is applied to various algorithms, but stronger in that the validation technique is also standardised, meaning that algorithms are not only validated on the same data, but that they do so in the same way and are compared in a uniform way as well. However, unlike with standardised datasets, there is always a doubt that the methods are being implemented fairly. That is, with the complexity of methods being presented today, there are innumerable parameters that would need to be set in a manner that is both cognisant of the problem domain and the algorithm itself. Thus, for all but a few very well-understood algorithms with open implementations, the team that is comparing the algorithms lacks some of the information of the original algorithm designers which can lead to sub-optimal versions of the algorithms being compared. This can at least partially be addressed through properly constructing shared code bases and using common tools and libraries. The capability of implementing a paper is at least a minimum metric of reproducibility.

However, the authors of a comparison paper may also (consciously or subconsciously) have a preferred method that receives additional attention which could subtly bias the validation results. Even if the exact code for a paper is provided, the variability in validation methodology across papers (as we've seen in Section 3) still provides an opportunity for a differential amount of effort to be applied in comparison of multiple frameworks.

4.5. Challenges

Challenges may address the issues with standardised open datasets and comparison papers at the same time. A *challenge* is a time-limited competition in which independent groups submit models (or results on pre-identified 'Testing' data) which are evaluated by the challenge organisers. Often, these challenges feature a *leaderboard* which ranks submissions according to one or more M_V. The idea of a challenge builds on that of standardised open datasets and comparison papers in that (at least) the 'Testing' data and validation procedure as a whole is standardised and uniformly applied across a variety of papers. In some ways, challenges combine the benefits of standardised open datasets as well as recommended validation approaches as they often involve teams who collect the data and design the experiment explicitly to cover a variety of metrics which are of clinical interest. The crucial difference between a challenge and a comparison paper is the separation of the group that constructs the algorithm, who are incentivised to create as high-performing an algorithm as possible given the particular data, and the group that performs the validation, who are incentivised to create a validation scheme that is both descriptive (covers a wide array of applications and metrics) and discriminative (i.e. distinguishes the performance of different algorithms).

A recent paper [41] explaining the Biomedical Image Analysis Challenges (BIAS) initiative emphasised the importance of challenges specifically in the MIC community as benchmarks to evaluate and reproduce a large number of algorithms. One of the issues discussed by this paper is the heterogeneity of challenge reporting, both in terms of

the datasets and objectives, but also in terms of how submissions are validated. In addition to the elements outlined earlier in this paper, BIAS includes validation information that is inherently specific to challenges, such as details on the ranking process, the challenge life-cycle, submission procedures, etc…These aspects do not reflect the validation of individual algorithms but are instead designed to make challenges more transparent and interpretable.

Maier-Hein et al. [39] has also demonstrated that quality control aspects such as ranking specifics, 'Testing' data composition, and reference determination have rendered the reproducibility and interpretation of challenge results problematic. Although the BIAS initiative can alleviate some of these issues by structured reporting, rendering the challenge results more interpretable by clearly specifying the framework for their interpretation, several theoretical and practical issues remain.

The design of challenges is difficult as it must not only provide a standard basis for comparison across multiple algorithms that may differ vastly in their technical specifics, but also address the underlying problems of human competitiveness and fallibility in a rigorous manner. For example, challenges tend to rely on the hold-out method as the validation technique, which is known to be less accurate than cross-validation techniques [8]. In addition, this burdens challenge designers to acquire sufficient data not only to evaluate the algorithms presented but also to give a meaningful opportunity for more data-hungry algorithms to compete. Maier-Hein et al. [39] found that the removal of a single 'Testing' case can change challenge rankings up to 67% of the time which reflects the instability of ranking procedures with respect to actual validation being performed. This is also important due to the uncertainty in the reference data. For example, the BRATS2014 segmentation challenge [46] had an inter-rater variability of 70-85% Dice with a few method approaching that level of accuracy. Maier-Hein et al. [39] found that for segmentation challenges in particular, the choice of observer (the clinician manually segmenting the imaging data) can result in changing the challenge ranking for between 15 and 62% of validation metrics. In over 60% of segmentation challenges, it is even unclear if multiple observers have segmented the data [39]. The lack of M_{RU} and/or M_{OU} information implies that, after a certain point, ranking results cannot be meaningfully interpreted as there is no mechanism to tell if two high-performing algorithms have truly different results.

Aside from reference data, challenges also struggle to eliminate data leakage. Often, the challenge organisers giving groups dedicated 'Testing' data (for 'hold-out' validation) to which the algorithms are applied, but this could still cause *human* data leakage when the participants change their algorithm based on the perceived results on said data. In order to completely eliminate this, some challenges are organised in which the participants give their models to the challenge organisers to be applied on standardised hardware to completely unseen 'Testing' datasets, although that is less common, requires the challenge organisers to have more extensive resources, and limits the scope of computing resources available to the participants.

None of this is to detract from the importance of challenges. By merely comparing multiple algorithms on the same datasets, challenges have significantly improved aspects of cross-algorithmic validation by controlling variability in experimental design. The contribution of this framework to challenges extends along the same lines as the previous section on open datasets (i.e. encouraging data augmentation features, reference uncertainty information, etc …) although the additional control available in challenges could lead to other opportunities. One possibility for improving challenges could be to use a cross-validation technique in order to evaluate how algorithms perform with limited access to data. This would however require more work for the organisers as they would have to create a structured environment for performing both the training and evaluation rather than only provide data.

5. Conclusion

This article has aimed to clarify the nature of medical information processing validation regarding both traditional methods and learning-based methods that have come to the forefront of MIC research since the advent of deep learning. Validation is an inherent part of any technical research, but is especially important in MIC as the difference between proper and improper validation can lead to problems in the integration of research into the clinic, a reduction in the trust a clinician would have for algorithmic assistance, or a negative impact on patient care.

This paper presents a framework for understanding the validation of medical information processing algorithms from a data-flow point of view, isolating the various components and considerations into distinct parts of a greater validation workflow. Although more complex than simpler work-flow based models previously proposed [26,27], this model is expressive enough to capture both traditional algorithms and machine-learning based algorithms into a singular framework which takes advantage of the similarities in validation philosophies between the two while highlighting crucial differences.

The overall result of this paper is a way of breaking down the process of validation in a way that facilitates comparison between different works in the literature. We demonstrated this through two literature reviews (one in subthalamic nucleus segmentation for deep brain stimulation interventions and another in cortical point localisation for transcranial magnetic stimulation) that are specifically catered towards a comparison of validation techniques. These literature reviews show how our framework can be easily put into place to rigorously analyse the validation aspects of different methods from the literature.

This framework, as well as helping to expose common issues in MIC algorithm validation, motivates methods for improving validation. Aside from providing a checklist of elements to include (and thus higher visibility of possible reporting issues and biases), such a framework motivates concrete mechanisms for community-wide improvements such as the creation of validation-critical literature reviews, published recommendations for validation approaches, standardised open datasets, comparison papers, and competitive challenges. These recommendations are not a panacea but they each represent possible methods for moving towards more robust algorithm validation.

Declaration of competing interest

The authors, John S.H. Baxter and Pierre Jannin, have no conflicts of interest to declare.

Acknowledgements

J.S.H. Baxter is supported by Institut national de la santé et de la recherche médicale (INSERM) and by the Institut des Neurosciences Cliniques de Rennes (INCR).

References

[1] Baxter JS, Bui QA, Maguet E, Croci S, Delmas A, Lefaucheur JP, Bredoux L, Jannin P. Automatic cortical target point localisation in mri for transcranial magnetic stimulation via a multi-resolution convolutional neural network. Int J Comput Assist Radiol Surg 2021:1–11.
[2] Baxter JS, Gibson E, Eagleson R, Peters TM. The semiotics of medical image segmentation. Med Image Anal 2018;44:54–71.
[3] Baxter JS, Jannin P. Bias in machine learning for computer-assisted surgery and medical image processing. 2022.
[4] Baxter JS, Jannin P. Combining simple interactivity and machine learning: a separable deep learning approach to subthalamic nucleus localization and segmentation in mri for deep brain stimulation surgical planning. J Med Imag 2022;9:045001.
[5] Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. J Mach Learn Res 2004;5:1089–105.
[6] Berrar D, Bradbury I, Dubitzky W. Avoiding model selection bias in small-sample genomic datasets. Bioinformatics 2006;22:1245–50.

[7] Blazis SP, Dickerscheid DB, Linsen PV, Jarnalo COM. Effect of ct reconstruction settings on the performance of a deep learning based lung nodule cad system. Eur J Radiol 2021;136:109526.
[8] Blum A, Kalai A, Langford J. Beating the hold-out: bounds for k-fold and progressive cross-validation. In: Proceedings of the twelfth annual conference on Computational learning theory; 1999. p. 203–8.
[9] Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, Jay M, Das R. A computational approach to early sepsis detection. Comput Biol Med 2016; 74:69–73.
[10] Cicuttini F, Forbes A, Morris K, Darling S, Bailey M, Stuckey S. Gender differences in knee cartilage volume as measured by magnetic resonance imaging. Osteoarthritis Cartilage 1999;7:265–71.
[11] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10:1895–923.
[12] Eklund A, Nichols TE, Knutsson H. Cluster failure: why fmri inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci USA 2016;113:7900–5.
[13] Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing 2018;321:321–31.
[14] Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88–94.
[15] Gach HM, Tanase C, Boada F. 2d & 3d shepp-logan phantom standards for mri. In: 2008 19th international conference on systems engineering. IEEE; 2008. p. 521–6.
[16] Gibson E, Li W, Sudre C, Fidon L, Shakir DI, Wang G, Eaton-Rosen Z, Gray R, Doel T, Hu Y, et al. Niftynet: a deep-learning platform for medical imaging. Comput Methods Progr Biomed 2018;158:113–22.
[17] Glatard T, Lartizien C, Gibaud B, Da Silva RF, Forestier G, Cervenansky F, Alessandrini M, Benoit-Cattin H, Bernard O, Camarasu-Pop S, et al. A virtual imaging platform for multi-modality medical image simulation. IEEE Trans Med Imag 2012;32:110–8.
[18] Gugino LD, Romero JR, Aglio L, Titone D, Ramirez M, Pascual-Leone A, Grimson E, Weisenfeld N, Kikinis R, Shenton ME. Transcranial magnetic stimulation coregistered with mri: a comparison of a guided versus blind stimulation technique and its effect on evoked compound muscle action potentials. Clin Neurophysiol 2001;112:1781–92.
[19] Hamarneh G, Jassi P. Vascusynth: simulating vascular trees for generating volumetric image data with ground-truth segmentation and tree analysis. Comput Med Imag Graph 2010;34:605–16.
[20] Han C, Hayashi H, Rundo L, Araki R, Shimoda W, Muramatsu S, Furukawa Y, Mauri G, Nakayama H. Gan-based synthetic brain mr image generation. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE; 2018. p. 734–8.
[21] Heckel F, Schwier M, Peitgen HO. Object-oriented application development with mevislab and python. Informatik 2009–Im Focus das Leben. 2009.
[22] Herwig U, Padberg F, Unger J, Spitzer M, Schönfeldt-Lecuona C. Transcranial magnetic stimulation in therapy studies: examination of the reliability of "standard" coil positioning by neuronavigation. Biol Psychiatr 2001;50:58–61.
[23] Holdcroft A. Gender bias in research: how does it affect evidence based medicine?. 2007.
[24] Hutson M. Artificial intelligence faces reproducibility crisis. Science 2018;359: 725–6. https://doi.org/10.1126/science.359.6377.725.
[25] Jannin P, Fitzpatrick JM, Hawkes D, Pennec X, Shahidi R, Vannier M. Validation of medical image processing in image-guided therapy. IEEE Trans Med Imag 2002;21: 1445–9.
[26] Jannin P, Grova C, Maurer CR. Model for defining and reporting reference-based validation protocols in medical image processing. Int J Comput Assist Radiol Surg 2006;1:63–73.
[27] Jannin P, Korb W. Assessment of image-guided interventions. In: Image-guided interventions. Springer; 2008. p. 531–49.
[28] Julkunen P, Säisänen L, Danner N, Niskanen E, Hukkanen T, Mervaala E, Könönen M. Comparison of navigated and non-navigated transcranial magnetic stimulation for motor cortex mapping, motor threshold and motor evoked potentials. Neuroimage 2009;44:790–5.
[29] Kapur T, Pieper S, Fedorov A, Fillion-Robin JC, Halle M, O'Donnell L, Lasso A, Ungi T, Pinter C, Finet J, et al. In: Increasing the impact of medical image computing using community-based open-access hackathons: the na-mic and 3d slicer experience; 2016.
[30] Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data 2012;6:1–21.
[31] Kikinis R, Pieper SD, Vosburgh KG. 3d slicer: a platform for subject-specific image analysis, visualization, and clinical support. In: Intraoperative imaging and image-guided therapy. Springer; 2014. p. 277–89.
[32] Kim J, Duchin Y, Kim H, Vitek J, Harel N, Sapiro G. Robust prediction of clinical deep brain stimulation target structures via the estimation of influential high-field mr atlases. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015. p. 587–94.
[33] Kim J, Duchin Y, Sapiro G, Vitek J, Harel N. Clinical deep brain stimulation region prediction using regression forests from high-field mri. In: 2015 IEEE international conference on image processing (ICIP). IEEE; 2015. p. 2480–4.
[34] Kim J, Duchin Y, Shamir RR, Patriat R, Vitek J, Harel N, Sapiro G. Automatic localization of the subthalamic nucleus on patient-specific clinical mri by incorporating 7 t mri and machine learning: application in deep brain stimulation. Hum Brain Mapp 2019;40:679–98.
[35] Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci USA 2020;117:12592–4.
[36] Li B, Jiang C, Li L, Zhang J, Meng D. Automated segmentation and reconstruction of the subthalamic nucleus in p arkinson's disease patients. Neuromodulation: Technology at the Neural Interface 2016;19:13–9.
[37] Luders E, Narr KL, Thompson PM, Rex DE, Woods RP, DeLuca H, Jancke L, Toga AW. Gender effects on cortical thickness and the influence of scaling. Hum Brain Mapp 2006;27:314–24.
[38] Lukacs PM, Burnham KP, Anderson DR. Model selection bias and freedman's paradox. Ann Inst Stat Math 2010;62:117.
[39] Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, Scholz P, Arbel T, Bogunovic H, Bradley AP, Carass A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 2018;9:1–13.
[40] Maier-Hein L, Reinke A, Christodoulou E, Glocker B, Godau P, Isensee F, Kleesiek J, Kozubek M, Reyes M, Riegler MA, et al. Metrics reloaded: pitfalls and recommendations for image analysis validation. 2022. arXiv preprint arXiv: 2206.01653.
[41] Maier-Hein L, Reinke A, Kozubek M, Martel AL, Arbel T, Eisenmann M, Hanbury A, Jannin P, Müller H, Onogur S, et al. Bias: transparent reporting of biomedical image analysis challenges. Med Image Anal 2020;66:101796.
[42] Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, Simuni T, Coffey C, Kieburtz K, Flagg E, Chowdhury S, et al. The Parkinson progression marker initiative (ppmi). Prog Neurobiol 2011;95:629–35.
[43] Martin PR, Cool DW, Romagnoli C, Fenster A, Ward AD. Magnetic resonance imaging-targeted, 3d transrectal ultrasound-guided fusion biopsy for prostate cancer: quantifying the impact of needle delivery error on diagnosis. Med Phys 2014;41:073504.
[44] McCormick MM, Liu X, Ibanez L, Jomier J, Marion C. Itk: enabling reproducible research and open science. Front Neuroinf 2014;8:13.
[45] McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021;13:eabb1655.
[46] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans Med Imag 2014;34:1993–2024.
[47] Milletari F, Ahmadi SA, Kroll C, Plate A, Rozanski V, Maiostre J, Levin J, Dietrich O, Ertl-Wagner B, Bötzel K, et al. Hough-cnn: deep learning for segmentation of deep brain regions in mri and ultrasound. Comput Vis Image Understand 2017;164:92–102.
[48] Moschidis E, Graham J. A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction. In: 2010 IEEE international symposium on biomedical imaging: from nano to macro. IEEE; 2010. p. 928–31.
[49] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). Alzheimer's Dementia 2005;1: 55–66.
[50] Mylius V, Ayache S, Ahdab R, Farhat W, Zouari H, Belke M, Brugières P, Wehrmann E, Krakow K, Timmesfeld N, et al. Definition of dlpfc and m1 according to anatomical landmarks for navigated brain stimulation: inter-rater reliability, accuracy, and influence of gender and age. Neuroimage 2013;78:224–32.
[51] Naganathan V, Macgregor A, Snieder H, Nguyen T, Spector T, Sambrook P. Gender differences in the genetic factors responsible for variation in bone density and ultrasound. J Bone Miner Res 2002;17:725–33.
[52] Nakkiran P, Kaplun G, Bansal Y, Yang T, Barak B, Sutskever I. Deep double descent: where bigger models and more data hurt. 2019. arXiv preprint arXiv:1912.02292.
[53] Oh JH, Yang Y, El Naqa I. Adaptive learning for relevance feedback: application to digital mammography. Med Phys 2010;37:4432–44.
[54] Olszowy W, Aston J, Rua C, Williams GB. Accurate autocorrelation modeling substantially improves fmri reliability. Nat Commun 2019;10:1–11.
[55] Peralta M, Jannin P, Baxter JS. Machine learning in deep brain stimulation: a systematic review. Artif Intell Med 2021:102198.
[56] Pesteie M, Abolmaesumi P, Rohling RN. Adaptive augmentation of medical data using independently conditional variational auto-encoders. IEEE Trans Med Imag 2019;38:2807–20.
[57] Polanski WH, Zolal A, Sitoci-Ficici KH, Hiepe P, Schackert G, Sobottka SB. Comparison of automatic segmentation algorithms for the subthalamic nucleus. Stereotact Funct Neurosurg 2020;98:256–62.
[58] Reijonen J, Könönen M, Tuunanen P, Määttä S, Julkunen P. Atlas-informed computational processing pipeline for individual targeting of brain areas for therapeutic navigated transcranial magnetic stimulation. Clin Neurophysiol 2021; 132:1612–21.
[59] Reinacher PC, Várkuti B, Krüger MT, Piroth T, Egger K, Roelz R, Coenen VA. Automatic segmentation of the subthalamic nucleus: a viable option to support planning and visualization of patient-specific targeting in deep brain stimulation. Operative Neurosurgery 2019;17:497–502.
[60] Rusjan PM, Barr MS, Farzan F, Arenovich T, Maller JJ, Fitzgerald PB, Daskalakis ZJ. Optimal transcranial magnetic stimulation coil placement for targeting the dorsolateral prefrontal cortex using novel magnetic resonance image-guided neuronavigation. Hum Brain Mapp 2010;31:1634–56.
[61] Russell TG, Jones AF. Implications of regulatory requirements for smartphones, gaming consoles and other devices. J Physiotherapy 2011;57(1);5-7.
[62] Samala RK, Chan HP, Hadjiiski L, Koneru S. Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. In: Medical imaging 2020: computer-aided diagnosis. International Society for Optics and Photonics; 2020, 1131416.
[63] Shamir RR, Duchin Y, Kim J, Patriat R, Marmor O, Bergman H, Vitek JL, Sapiro G, Bick A, Eliahou R, et al. Microelectrode recordings validate the clinical

visualization of subthalamic-nucleus based on 7t magnetic resonance imaging and machine learning for deep brain stimulation surgery. Neurosurgery 2019;84:749–57.
[64] Shepp LA, Logan BF. The fourier reconstruction of a head section. IEEE Trans Nucl Sci 1974;21:21–43.
[65] Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ open respiratory research 2017;4:e000234.
[66] Smith D, Welch E. Non-sparse phantom for compressed sensing mri reconstruction. In: International society for magnetic resonance in medicine 19th scientific meeting-ISMRM; 2011. p. 2845.
[67] Sparing R, Buelte D, Meister IG, Pauš T, Fink GR. Transcranial magnetic stimulation and the challenge of coil placement: a comparison of conventional and stereotaxic neuronavigational strategies. Hum Brain Mapp 2008;29:82–96.
[68] Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. NPJ digital medicine 2019;2:1–3.
[69] Sun Y, Yuan J, Rajchl M, Qiu W, Romagnoli C, Fenster A. Efficient convex optimization approach to 3d non-rigid mr-trus registration. In: International conference on medical image computing and computer-assisted intervention. Springer; 2013. p. 195–202.
[70] Tahhan AS, Vaduganathan M, Greene SJ, Alrohaibani A, Raad M, Gafeer M, Mehran R, Fonarow GC, Douglas PS, Bhatt DL, et al. Enrollment of older patients, women, and racial/ethnic minority groups in contemporary acute coronary syndrome clinical trials: a systematic review. JAMA cardiology 2020;5:714–22.
[71] Top A, Hamarneh G, Abugharbieh R. Active learning for interactive 3d image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2011. p. 603–10.
[72] Tran NT, Tran VH, Nguyen NB, Nguyen TK, Cheung NM. On data augmentation for gan training. IEEE Trans Image Process 2021;30:1882–97.
[73] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys 2018;102:1143–58.
[74] Trzasko J, Manduca A. Highly undersampled magnetic resonance image reconstruction via homotopic \ell_{0} -minimization. IEEE Trans Med Imag 2008;28:106–21.
[75] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14:e0224365.
[76] Vannier MW, Staab EV, Clarke LC. Medical image archives—present and future. In: CARS 2002 computer assisted radiology and surgery. Springer; 2002. p. 565–70.
[77] Varga I, Bakstein E, Gilmore G, Novak D. Image-based subthalamic nucleus segmentation for deep brain surgery with electrophysiology aided refinement. In: Multimodal learning for clinical decision support and clinical image-based procedures. Springer; 2020. p. 34–43.
[78] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinf 2006;7:1–8.
[79] Wallace DR, Fujii RU. Software verification and validation: an overview. Ieee Software 1989;6:10–7.
[80] Wiles AD, Likholyot A, Frantz DD, Peters TM. A statistical model for point-based target registration error with anisotropic fiducial localizer error. IEEE Trans Med Imag 2008;27:378–90.
[81] Xu J, Kobayashi S, Yamaguchi S, Iijima Ki, Okada K, Yamashita K. Gender effects on age-related changes in brain structure. Am J Neuroradiol 2000;21:112–8.
[82] Yagis E, De Herrera AGS, Citi L. Generalization performance of deep learning models in neurodegenerative disease classification. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2019. p. 1692–8.
[83] Zhou B, Chen L, Wang Z. Interactive deep editing framework for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2019. p. 329–37.
[84] Zhou T, Li L, Bredell G, Li J, Unkelbach J, Konukoglu E. Volumetric memory network for interactive medical image segmentation. Med Image Anal 2023;83:102599.
[85] Zosso D, Noirhomme Q, Davare M, Macq B, Olivier E, Thiran J, De Craene M. Normalization of transcranial magnetic stimulation points by means of atlas registration. In: 2006 14th European signal processing conference. IEEE; 2006. p. 1–5.