Value and limitations of intracranial recordings for validating electric field modeling for transcranial brain stimulation

Puonti, Oula; Saturnino, Guilherme B.; Madsen, Kristoffer H.; Thielscher, Axel

Published in:
NeuroImage

Link to article, DOI:
10.1016/j.neuroimage.2019.116431

Publication date:
2020

Document Version
Publisher’s PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Value and limitations of intracranial recordings for validating electric field modeling for transcranial brain stimulation

Oula Puonti a,b,1, Guilherme B. Saturnino a,b,1, Kristoffer H. Madsen a,c, Axel Thielscher a,b,*

a Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Hvidovre, Denmark
b Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark
c Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

ARTICLE INFO

Keywords:
Transcranial brain stimulation
TDCS
TACS
Volume conductor model
Errors-in-variables regression
Bayesian regression

ABSTRACT

Comparing electric field simulations from individualized head models against in-vivo intra-cranial recordings is considered the gold standard for direct validation of computational field modeling for transcranial brain stimulation and brain mapping techniques such as electro- and magnetoencephalography. The measurements also help to improve simulation accuracy by pinning down the factors having the largest influence on the simulations. Here we compare field simulations from four different automated pipelines against intracranial voltage recordings in an existing dataset of 14 epilepsy patients. We show that modeling differences in the pipelines lead to notable differences in the simulated electric field distributions that are often large enough to change the conclusions regarding the dose distribution and strength in the brain. Specifically, differences in the automatic segmentations of the head anatomy from structural magnetic resonance images are a major factor contributing to the observed field differences. However, the differences in the simulated fields are not reflected in the comparison between the simulations and intra-cranial measurements. This apparent mismatch is partly explained by the noisiness of the intra-cranial measurements, which renders comparisons between the methods inconclusive. We further demonstrate that a standard regression analysis, which ignores uncertainties in the simulations, leads to a strong bias in the estimated linear relationship between simulated and measured fields. Ignoring this bias leads to the incorrect conclusion that the models systematically misestimate the field strength in the brain. We propose a new Bayesian regression analysis of the data that yields unbiased parameter estimates, along with their uncertainties, and gives further insights to the fit between simulations and measurements. Specifically, the unbiased results give only weak support for systematic misestimations of the fields by the models.

1. Introduction

Modeling the current flow distribution in the brain is in the core of many neuroimaging and functional brain mapping techniques. The currents-of-interest can be either externally induced by transcranial brain stimulation (TBS) methods, such as transcranial electrical stimulation (TES) or transcranial magnetic stimulation (TMS) or can result from neuronal activity in which case they can be measured using potential differences on the scalp (EEG) or by recording the produced magnetic fields (MEG). In TBS, the individual anatomy has a large, and often counter-intuitive, impact in shaping the current flow inside the cranium (Bungert et al., 2017; Datta et al., 2009; Laakso et al., 2015; Miranda et al., 2013; Opitz et al., 2015). Similarly, the signal measured in EEG and MEG is dependent on the complex geometry of the head (Cho et al., 2015; Dannhauer et al., 2011; Stenroos et al., 2014). These findings have prompted a shift away from simplified anatomical models (Ravazzani et al., 1996) towards individualized head models based on structural magnetic resonance (MR) scans. Individualized modeling holds great promise particularly for TES, where several studies revealed a large inter-subject variability of the physiological stimulation effects (e.g., López-Alonso et al., 2014; Wiethoff et al., 2014). Part of this variability is likely explained by dosing differences due to anatomical variation. Individualized modeling enables dose control and can be used to systematically improve spatial targeting by automated tailoring of the...
electrode positions and injected currents, ensuring that the highest field strengths are contained to the region-of-interest (Dmochowski et al., 2011; Saturnino et al., 2019a). This opens the door for personalized treatment approaches in a variety of brain disorders ranging from major depressive disorder (Cifcsák et al., 2018) to motor rehabilitation after stroke (Datta et al., 2011; Minjoli et al., 2017).

Although practically relevant results have been obtained from simulation studies, which support the usefulness of individualized head models, one of the key challenges is the direct in-vivo validation of the electric field simulations in the human brain. The modeling process includes uncertainties, mainly related to the segmentation of the anatomy (Nielsen et al., 2018) and spatial tissue conductivities (Saturnino et al., 2019b), which propagate onto the estimated fields. Optimally, one would use in-vivo field measurements in the brain to gauge the accuracy of the simulations. In practice this is, however, difficult: In-vivo measurements of the electric fields are experimentally very demanding and susceptible to measurement errors, creating unwanted uncertainty in the data that is supposed to be used as ground truth for validating the simulations. To-date, we are aware of three studies where head model validation using direct in-vivo intra-cranial measurements of the electric fields in humans is attempted. The first one by (Opitz et al., 2016) is focused on validating the assumption that the head acts as an ohmic conductor. The second one by (Huang et al., 2017) reports TES-induced voltage measurements on ten epilepsy patients with intra-cranial electrodes. The authors use the voltage measurements for assessing the correlations between the simulated and measured voltage differences and for calibrating the tissue conductivities of the individual head models. In similar vein (Opitz et al., 2018), compare simulated and measured fields in two epilepsy patients reporting slightly lower correlations compared to those in (Huang et al., 2017). The difference is probably explained by differences in the experimental procedures, as the recording and TES electrodes were quite close to each other and to nearby skull defects in the study of (Opitz et al., 2018), suggesting that discrepancies between real and modeled positions and anatomy might have had stronger effects on the field comparisons than in (Huang et al., 2017). (Huang et al., 2017) found that models based on “standard” literature values for the ohmic conductivities systematically overestimated the recorded voltage differences. On the other hand (Opitz et al., 2018), found underestimated fields in one of the studied patients, while the calculated e-fields were too high in the second patient. Recently (Göksu et al., 2018), demonstrated a novel non-invasive approach to reconstruct TES induced current densities in the brain from MR images of the current-generated magnetic fields (magnetic resonance current density imaging, MRCDI). They presented initial results on five subjects showing good agreement between simulated and measured current densities, with a moderate but systematic underestimation of the current densities by the models based on “standard” ohmic conductivities. Non-invasive measurements would be the preferred approach, not only due to ethical aspects relating to invasive studies, but also because invasive measurements change the volume conduction properties of the head, thus introducing additional modeling complexities. MRCDI is a promising step to the correct direction but needs further development before it can be applied for head model validation.

In this article, we compare four different automated methods for end-to-end electric field simulations starting from a structural MR scan, followed by segmentation of the anatomy and generation of a finite element (FEM) mesh, and finally calculating the electric field distribution in the brain for a given stimulation protocol. We reproduce and extend the analysis presented in (Huang et al., 2017) and (Huang et al., 2019) using a freely available data set from (Huang et al., 2017). Specifically, we set out to demonstrate four points: first, differences in the modeling pipelines, such as the choice of segmentation and FEM approaches, often result in clear differences in the electric field simulations that are in the range of what is considered physiologically relevant. Second, the field differences depend on the size of the differences between the automatic head segmentations. Third, the field differences are not reflected in the comparison between the simulated and measured fields, which can be partly explained by the limitations of the validation data. Fourth, applying a standard linear regression analysis to compare the simulations and measurements leads to a biased estimate of the linear relationship between the two. In contrast, a more elaborate Bayesian regression analysis overcomes this problem, and allows for quantifying the uncertainty in the parameter estimates, which helps the interpretation of the fits between the simulations and measurements.

The results highlight the difficulty of validating the simulations, even when direct measurements are available, and point to a need for a more careful analysis of the available data and for adopting a strategic approach to future measurement studies in order to reach conclusive validations.

2. Material and methods

The data set consists of 14 epilepsy patients with intracranial EEG electrodes planted for surgical evaluation (Huang et al., 2017). For each subject there exists a T1-weighted MR scan, a manually corrected segmentation of the main tissue classes (white matter - WM, gray matter - GM, cerebro-spinal fluid - CSF, skull and scalp) along with annotations of the extra-cranial stimulation electrodes, subgaleal electrodes, intracranial electrode strip, and the surgical brain. The locations, in MNI and voxel coordinates, and measured voltages from the intracranial electrodes are provided as a text file. In general, the stimulation electrodes are placed medially over the frontal and occipital poles, with some exceptions, and transcranial alternating current stimulation (TACS) is performed with a 1 mA baseline peak at 1 Hz (Huang et al., 2017). Based on the manually corrected segmentations, we modeled the electrodes as 2 × 2 cm squares with 3 mm gel and rubber layers and determined their locations manually in each subject.

We compare two different software tools for generating individualized head models and simulating the electric fields induced by TES: SimNIBS 2.1 (Saturnino et al., 2019) and ROAST v2.7 (Huang et al., 2019). SimNIBS 2.1 offers three alternative approaches for generating the anatomical head models, which we consider individually, giving in total four methods to compare. For completeness, we will next briefly describe each of the approaches.

2.1. Head model generation

- The default pipeline for head model generation in SimNIBS 2.1, called headreco (Nielsen et al., 2018), uses the segmentation routine from SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) (Ashburner and Friston, 2005), combined with an extended anatomical atlas (Huang et al., 2013), to generate a tissue segmentation from a set of possibly multi-contrast MR scans. After the initial segmentation, the tissue masks are cleaned using simple morphological operations to reduce noise and ensure that the tissues are contained within each other. Next, surfaces, represented as triangular elements, are extracted from the voxel segmentations. As a last step, the FEM mesh is generated by filling in the space between the surfaces with tetrahedra. Note, that due to the chosen meshing approach, which first generates surfaces, the tissue classes need to be nested which is ensured by the clean-up step after the initial segmentation. For further details we refer the reader to (Nielsen et al., 2018).
- The headreco pipeline supports detailed cortical surface reconstructions, which are generated using the computational anatomy toolbox (CAT12, http://www.neuro.uni-jena.de/cat/) implemented in SPM12. The cortical surfaces from CAT12 replace the ones generated from the voxel segmentations in the standard headreco pipeline, while other parts of the pipeline remain the same. We denote this approach headreco + CAT.
- The predecessor of headreco in SimNIBS, called mri2mesh (Windhöff et al., 2013), combines the cortical surfaces and subcortical segmentation generated by FreeSurfer (https://surfer.nmr.mgh.harvard.edu/) (Fischl et al., 2002), with extra-cerebral tissue
porting in (Huang et al., 2017). Similar to headreco with CAT12, the cortical surfaces from FreeSurfer are combined with surfaces created from voxel segmentations in the other tissues, and the final FEM mesh is obtained by filling in tetrahedra.

Similar to the standard version of headreco, the ROAST v2.7 toolbox (https://www.parralab.org/roast/) (Huang et al., 2019) generates the tissue segmentations using SPM12, with the same extended anatomical atlas, and applies morphological operations to clean the segmentations. The main difference between the methods is in the post-processing and FEM meshing approach: whereas headreco first creates surfaces, ROAST generates a tetrahedral volume mesh directly from the voxel segmentation using CGAL (Fabri and Teillaud, 2011) called through the iso2mesh (http://iso2mesh.sourceforge.net/cgi-bin/index.cgi) (Fang and Boas, 2009) toolbox. The restriction of nested tissue classes is thus relaxed, and anatomical details can potentially be better captured if the initial volume segmentation is accurate. However, the reconstructions of the tissue boundaries can be less accurate as the volume segmentation is meshed directly.

Data and code availability statement: The data set is freely available for download after registration at https://doi.org/10.6080/K0XW4GQ1. The software tools used for the analysis are distributed freely as open-source packages. The code for running the data analysis, described in the Analyses-section below, is included in the supplementary material.

2.2. Simulating the electric fields

The electric field calculations are performed in SimNIBS 2.1 for the head models generated with its pipelines (headreco, headreco + CAT and mri2mesh) and in ROAST v2.7 for the head models generated with ROAST. The current flowing through the electrodes is set to 1 mA, and the polarity adjusted to fit the recordings from (Huang et al., 2017) so that the direction of current flow is consistent with the measured data. Tissue and electrode conductivities were set to the literature values reported in (Huang et al., 2017).

Both SimNIBS 2.1 and ROAST v2.7 use the GetDP (Geuzaine, 2007) software to calculate electric potentials using the FEM method with first order tetrahedral elements. However, the post-processing of the simulations differs: ROAST uses GetDP to calculate the electric fields in each mesh node, while SimNIBS has native post-processing functions calculating the electric field for each mesh tetrahedra. The post-processing in SimNIBS is more consistent with the mathematical formulation of the Finite Element Method, where gradients are defined element-wise instead of node-wise (Zienkiewicz et al., 2013), and yields more physically plausible results, as the electric field values are discontinuous across tissue interfaces (Geselowitz, 1967). When interpolating or gridding results, SimNIBS uses the original mesh grid, keeping geometric consistency, while ROAST uses the TriScatteredInterp function in MATLAB (MathWorks, 2019), which does not preserve the original mesh, and instead creates a new Delaunay triangulation where the gridding is performed. Thus, the electric field values interpolated in ROAST do not observe tissue boundaries, as they do in SimNIBS.

3. Analyses

We performed two sets of analyses: the first one to quantify the differences in anatomical segmentation accuracy along with the differences in the simulated electric field distributions between the methods, and the second one to relate the electric field simulations to the measured potential differences in the intra-cranial electrodes. All the pipelines were run with default settings with the following exceptions:

- Both headreco and headreco + CAT were run with the –d no-conform option to avoid resampling of the input scans.
- For P04 in headreco + CAT, we set the vertex density (-v option) to 1.5 Nodes/mm²
- For P014 in mri2mesh, we set the number of vertices (–nvertices option) to 120000.
- For P010 the MR scan was resampled to 1 mm³ isotropic as ROAST v2.7 does not account for anisotropic scans resulting in erroneous electric field estimates by effectively changing the electric conductivities along the axis where the anisotropy occurs. In the meantime, this bug has been fixed in a newer version of ROAST (2.7.1).
- For P06, we inverted the “x” component of the electric field calculated with ROAST to account for the fact that ROAST does not correct for the “x” axis flipping indicated by the header in the NiTi image.

The changes to the vertex densities in P04 and P014 were made as, after running the head model pipeline, we found that the head meshes were missing volumes (WM in P04 and CSF in P014). In both cases, increasing mesh density made surface decoupling more accurate and thus solved the problems in meshing the surfaces. The average edge size, number of nodes and tetrahedra in the final meshed obtained with each method is shown in Table 1.

3.1. Variability in segmentations and electric fields

Assessing the anatomical segmentation accuracy of the four methods requires a ground truth segmentation to compare against. The manually corrected segmentations were created by first running the ROAST segmentation tool on the T1-weighted scans, then automatically correcting the output using a custom script, and finally correcting the remaining errors by hand (Huang et al., 2017). However, the data set is very challenging to segment due to the relatively low (clinical) MR scan quality and surgical interventions, and some of the manually corrected segmentations still have inaccuracies (see Fig. 3). As the segmentation procedure is based on ROAST, the manually corrected segmentations could also be biased towards the automated ROAST, headreco, and headreco + CAT segmentations, which all use SPM12 to segment the head tissues. To partially correct this issue, we generated a “consensus” head segmentation based on a multi-atlas approach using majority voting (Iglesias and Sabuncu, 2015). There, each of the segmentations obtained from the four automated pipelines, as well as the manually corrected segmentation, cast a single vote on the classification of each voxel, and the tissue with the most votes is selected. We can then compare the individual segmentations to the consensus using the Dice overlap score. The Dice score is defined as:

\[
\text{Dice}(C,A) = \frac{2|C \cap A|}{|C| + |A|}
\]

where \(C\) and \(A\) denote the consensus and the automated, or manually corrected, segmentation masks of a given tissue. It serves as an indication of the general segmentation differences between the four methods, such that we can see if a segmentation method consistently deviates from the consensus.

The differences between the simulated electric fields given by the methods were measured calculating the relative difference in the fields in Table 1

<table>
<thead>
<tr>
<th>Average edge length, number of nodes and number of tetrahedra across all meshes.</th>
<th>mri2mesh</th>
<th>headreco + CAT</th>
<th>headreco</th>
<th>ROAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average edge length (mm)</td>
<td>1.95</td>
<td>2.12</td>
<td>2.21</td>
<td>2.78</td>
</tr>
<tr>
<td>Number of nodes</td>
<td>0.72 × 10⁶</td>
<td>1.00 × 10⁶</td>
<td>0.74 × 10⁶</td>
<td>0.32 × 10⁶</td>
</tr>
<tr>
<td>Number of tetrahedra</td>
<td>4.03 × 10⁶</td>
<td>5.50 × 10⁶</td>
<td>4.02 × 10⁶</td>
<td>1.88 × 10⁶</td>
</tr>
</tbody>
</table>
Fig. 1. Dice scores computed for each of the pipelines, and the manually corrected segmentations, in WM, GM, CSF, skull and scalp. On each box, the line marks the median, the box extends to lower and upper quartiles, whiskers extend up to 1.5 times the interquartile range, and data points beyond that are marked as outliers. The higher the score, the more similar the segmentations are to the consensus.

Fig. 2. Example segmentations from a subject where all segmentations agree. From left to right, top to bottom: the input T1-weighted MRI scan, consensus segmentation, manually corrected segmentation, ROAST, headreco + CAT, headreco and mri2mesh. Note that mri2mesh does not segment the subcortical gray matter. The black lines close to the cortex in the manually corrected segmentations correspond to the intra-cranial electrode strips. Lowest row shows the differences in the norm of the electric field on the cortex.
GM for each pair of methods. The relative field difference, akin to the Dice score, is defined as:

$$e\text{Diff}(E_1, E_2) = \frac{2 ||E_1 - E_2||}{||E_1 + E_2||}.$$  

(2)

The differences were evaluated in the middle cortical surface obtained from the CAT12 segmentation. Electric field values were interpolated from the gray matter region of the FEM meshes to the cortical surface. If necessary, electric field values were extrapolated by taking the nearest gray matter neighbor.

### 3.2. Fitting intracranial measurements

In this analysis, we wanted to relate the intracranial voltage recordings to the field simulations. As the electric field was not measured, but rather the voltage relative to a reference contact, we calculated pairwise voltage differences between consecutive electrode contacts and divided them by the distance between contacts. This corresponds to a coarse estimate of the electric field component along the electrode axis. To provide an unbiased comparison, the same procedure was done with the simulations, where the simulated voltages were sampled in the contact locations in SimNIBS, by performing barycentric interpolation based on the electric potentials calculated in the mesh nodes, and in ROAST by interpolating the gridded voltage values.

Next, we fitted a standard linear model for each subject and method:

$$y_s = \beta_{s,m} x_{s,m} + \varepsilon_{s,m}$$

$$\varepsilon_{s,m} \sim N(0, \sigma_{s,m}).$$

or, equivalently

$$y_s \sim N(\beta_{s,m} x_{s,m}, \sigma_{s,m}).$$

(3)

Where $N(\mu, \sigma)$ denotes a normal distribution of mean $\mu$ and standard deviation $\sigma$. $x_{s,m}$ is a vector of the simulated potential differences for subject $s$ and method $m$, $y_s$ is a vector of the recorded potential differences in subject $s$, and the noise $\varepsilon_{s,m}$ for each subject and method is assumed to be drawn from a normal distribution with mean zero and standard deviation $\sigma_{s,m}$. For each subject and segmentation, we report the slope $\beta_{s,m}$, the coefficient of determination ($r^2$) and correlation ($\rho$). This serves as a re-analysis of the results by (Huang et al., 2019, 2017) in comparing simulations to intracranial electrode recordings.

We found that the standard slope estimates were correlated with the measured field strengths, indicating that subjects with low signal had a systematic bias towards underestimated regression fits (see Results section for details). To account for the bias, we performed a hierarchical Bayesian regression analysis, where the slope of each subject and method $\beta_{s,m}$ is drawn from an underlying distribution for the group-level slope of each method $\beta_m$. In addition, we adopted a Bayesian errors-in-variables model (Gull, 2013; Minka, 1999), which allows accounting for noise in the measurements as well as uncertainties in the simulations that arise from noise in the MR scans, uncertain electrical conductivity values and segmentation errors. In short, the regression model now becomes:

$$y_s \sim N(\beta_{s,m} x_{s,m}, \sigma_{s,m}).$$

(4)
\[ x_{m,n} \sim N(\beta_{m,n}, \sigma_{n}) \]  

(5)

where \( x_{m,n} \) is a vector of the unobserved “true” simulated values, and \( x_{m,n} \) is a vector of the observed simulated potential differences. Note that in this model, not only measurement noise is considered, but also uncertainties in the simulations. We further assume that the slopes are generated as:

\[ \beta_{m,n} \sim N(\beta_{m}, \sigma_{m}) \]  

(6)

where \( \beta_{m,n} \) is the unobserved hyperparameter for the group average slope of method \( m \), and the standard deviation \( \sigma_{m} \) captures the subject-level variation of the slopes. If the estimated subject-specific slopes are close to each other this variation will be small, whereas if they are far apart the variation will be large. We further need to define prior distributions on the noise parameters, the average slope \( \beta_{m} \), and the unobserved “true” simulations \( x_{m} \). These, along with the full modeling details, can be found in the supplementary material. The Bayesian analysis was performed using Stan (Carpenter et al., 2017), specifically PyStan (https://pystan.readthedocs.io/en/latest/) for Python interfacing. The Stan code for running the analysis is provided in Supplementary Material 1.

The benefit of adopting this type of Bayesian modeling is three-fold: first, as mentioned before, the noise and uncertainties in the measurements and simulations can both be estimated in a principled way. Second, we can study the posterior distributions of the slope for each subject and method to see which values are supported by the data given the model. Third, we can evaluate the group differences between the methods using the posterior predictive distribution of the slope for an unseen subject.

### 4. Results

#### 4.1. Variability in segmentations and electric fields

Here we aim to show that the different modeling pipelines yield different segmentations and electric field distributions, and that the differences between the simulations are large. Fig. 1 shows the Dice scores comparing the automated and manually corrected segmentations to the consensus segmentation in the five main head tissue classes: WM, GM, CSF, skull and scalp. The average Dice scores over all subjects and tissues, along with the standard deviations, for each method are: \( 0.757 \pm 0.097 \) (\textit{mri2mesh}), \( 0.908 \pm 0.027 \) (headreco + CAT), \( 0.933 \pm 0.040 \) (headreco), \( 0.836 \pm 0.137 \) (ROAST), and \( 0.903 \pm 0.078 \) (Manually Corrected). On average, the two headreco pipelines and the \textit{Manually Corrected} segmentations are closest to the consensus obtaining a Dice score above 0.9, followed by ROAST and finally \textit{mri2mesh}. As the consensus segmentation is based on votes from the different methods, there is a chance that it might be unfairly biased towards a subset of the segmentation approaches. In this study three of the five approaches originate from SimNIBS, which could lead to ROAST having lower Dice scores due to the consensus segmentation agreeing more with SimNIBS-based methods. However, as four of the five approaches, namely headreco, headreco + CAT, ROAST and \textit{Manually Corrected}, are based on SPM12 this is unlikely to be the case. To verify this, we additionally computed the Dice scores compared to a consensus segmentation where one of the SimNIBS methods (headreco) is excluded, see Supplementary Material 3. The Dice scores of ROAST and the \textit{Manually Corrected} segmentations even decrease slightly, implying that the Dice scores in Fig. 1 are not favoring the segmentation approaches implemented in SimNIBS.

To get a better understanding where the segmentation differences arise from, we picked the subject where the methods agreed most (Fig. 2), i.e., highest average Dice score over tissues, subjects, and methods, and the subject where the methods disagreed the most (Fig. 3), i.e., lowest average Dice score over tissues, subjects, and methods. Two additional subjects with second-best and second-worst agreements are shown in the Supplementary Material 3, Figs. S22 and S23.

The differences in the norm of the electric field in the cortex exceed in part 50%, and also the positions of the most strongly simulated brain areas (the “hot spots”) vary across methods. This would clearly change our interpretation of which brain areas get most strongly stimulated and how strong the stimulation effects might be. For example, in the case of the subject in Fig. 3, the e-field simulations based on \textit{mri2mesh} reach 0.2 V/m in the prefrontal cortex, but they hardly exceed 0.1 V/m in the simulations based on ROAST. While there is no consensus about the minimal field intensities that are required to cause reliable physiological stimulation effects, the data available so far from in-vitro and invasive recordings in animals suggest that fields stronger than 0.2 V/m are able to affect the neural activity under favorable conditions while fields below that might lack neural effects (Liu et al., 2018). Applying this threshold to the subject in Fig. 3, one would conclude that several brain areas got stimulated when considering the simulations based on \textit{mri2mesh}, but that stimulation was fully ineffective when considering the ROAST results. In addition, only the simulations based on \textit{mri2mesh} would indicate a stimulation of the temporal lobe, while this would not be the case when considering the other three simulation results.

It is also clear that the MR scans of the two subjects where the methods agree (Fig. 2 and S22), are of better quality in terms of contrast between tissues compared to the MR scans of the two subjects where the methods disagree (Fig. 3 and S23). Specifically, the gray-white matter contrast is higher in the T1w scans where the methods agree, whereas the T1w scans where the methods disagree seem to be contrast-enhanced, see e.g., the superior sagittal sinus posterior to the ventricles in Fig. 3 and S23. This shows that the uncertainties in the segmentations are directly related to input MR data quality, resulting in more disagreement between the methods when the contrast between tissues is poor. The poor agreement between \textit{mri2mesh} and the consensus is explained by the fact that \textit{mri2mesh} segments the subcortical GM and the whole cerebellum as WM, resulting in lower Dice scores for WM and GM. Furthermore, the extra-cerebral segmentations rely on a fairly simple method, which has been shown to be outperformed by the SPM12-based approaches (Nielsen et al., 2018). \textit{mri2mesh} also does not model the air pockets in the head, which might affect the fields estimates for some electrode montages located close to the sinuses. The largest difference between the two headreco pipelines and ROAST seems to be that the ROAST segmentation is generally less smooth in the sense that the tissue segmentations are not spatially continuous. This is apparent even in the two cases where the MR contrast is good (Fig. 2 and S22). The consensus segmentation allows for visualizing the systematic segmentation errors across all methods by studying the two consensus segmentations where the methods agree the most (Fig. 2 and S22). We observe that the sulcal CSF seems to be often segmented as GM, and that the skull is under-segmented when spongy bone is present. Example segmentations of all the subjects from the automated methods are provided in Supplementary Material 2.

Next, we study the differences in the simulated electric fields between the automated methods. Table 2 shows the average relative difference between the electric field simulations (eDiff, Eq. (2)) for each pair of methods. The results from headreco and headreco + CAT agree most, which is expected as the only difference between the two is in the GM surface reconstruction. \textit{mri2mesh} is the most different from all other methods, with differences in the range of upper 40%, which is likely due to two factors: first, the segmentation approach is different from the other methods, i.e., not based on SPM12, and second, \textit{mri2mesh} does not

<table>
<thead>
<tr>
<th>Mean electric field difference (eDiff)</th>
<th>ROAST</th>
<th>headreco</th>
<th>headreco + CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{mri2mesh}</td>
<td>49.0% ± 11.1%</td>
<td>34.1% ± 6.7%</td>
<td>34.1% ± 8.0%</td>
</tr>
<tr>
<td>headreco + CAT</td>
<td>36.6% ± 12.3%</td>
<td>21.4% ± 2.7%</td>
<td></td>
</tr>
<tr>
<td>headreco</td>
<td>34.3% ± 13.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
model the neck resulting in differences in electrode placement in those subjects with an electrode in the neck. ROAST has differences around 25% from the headreco methods, all of which share large parts of the segmentation algorithm, as they are based on SPM12. However, post-processing of the segmentations and the electric fields differs and is likely to cause most of the observed differences in the electric fields.

To link the segmentation differences to e-field differences, we calculated the average Dice scores over the tissues between the segmentations from each pair of methods and split the subjects to low and high Dice score groups based on the median. We then plotted the relative e-field differences (eDiff, Eq. (2)) in both groups for each pair of methods, which are shown in Fig. 4. In all comparisons the higher Dice score group, i.e., above median, has a lower relative e-field difference than the lower Dice score group, indicating that when the segmentations agree so do the field simulations. This effect can be seen also in the small difference between the high and low Dice score groups between headreco and headreco + CAT as the segmentations from both approaches are very similar to each other.

We visually explored the simulation results to get a qualitative overview of the typical segmentation differences that cause the e-field differences. Some informative examples are shown in Fig. 5. In general, we see that the amount of CSF has a large effect on the simulated electric fields likely due to shunting effects. The first row in Fig. 5 shows that if the amount of CSF is less, the simulated fields in the cortex can be much higher as the current does not redistribute through the highly conducting CSF. Thus, accurate segmentation of the GM sulci also becomes important for locally accurate field modeling. The second row shows a similar effect, where the skull is mislabeled either as CSF (left) or scalp (right). The final row in Fig. 5 shows spurious islands of GM voxels in the ROAST segmentation, which can lead to extremely high field estimates in GM as these voxels are close to skull and surrounded by CSF. We note that segmenting out the CSF on this data set is challenging as no T2-weighted (T2w) scan is provided. As the skull-CSF border is highly visible in T2w scans, they typically contribute to an accurate placement of the skull-CSF border (Nielsen et al., 2018).

In Fig. 6, we show the norm of the electric field in WM, GM and CSF for both ROAST and headreco + CAT in subject P03. The effect of the different electric field post-processing schemes between SimNIBS and ROAST is quite striking: the interpolated field in ROAST is blurred, making the WM-GM border invisible and causing the large electric field estimates in the skull to bleed into CSF and to a lesser extent into GM. This effect makes the electric field estimates for CSF in ROAST clearly overestimated, as the fields in CSF are lower due to its high electrical conductivity.

4.2. Fitting intracranial measurements

4.2.1. Standard regression analysis

We first present the results from a standard regression analysis to reproduce the comparison of the methods from (Huang et al., 2019). Fig. 7 shows the coefficient of determination ($r^2$) and the slope ($\beta_s$; $m$) of the standard linear regression (Eq. (3)) for each subject and method, and Table 3 shows the mean and standard deviation of both quantities, along with the correlation ($\rho$), across all subjects. Assessing Fig. 7 qualitatively, it seems that all methods perform approximately equal in predicting the recordings in terms of both the $r^2$ and the slope. Testing for differences using a one-way repeated measures ANOVA revealed no statistically significant differences in the $r^2$ values between the methods ($p = 0.065$) but did so for the slope values ($p = 0.021$). However, pairwise post hoc comparisons between the slope estimates of methods did not reveal any
differences using Tukey’s Honest Significant Difference test. When not considering multiple comparisons in the post hoc tests, we find trends of differences between the slope estimates of headreco and headreco + CAT (p = 0.009), headreco and ROAST (p = 0.017), and headreco + CAT and mri2mesh (p = 0.027). The range of explained variance (r^2) seems to be large over the subjects, where in some subjects (P04 and P07) the modeled fields explain the measurements well, while in others (P06 and P09) the prediction is poor. These results are in line with the ones reported in (Huang et al., 2019, 2017; Opitz et al., 2018). In addition, we also observe that all methods tend to overestimate the measured potential differences as reported by (Huang et al., 2019), but that the correlations are similar for all methods and close to the ones reported for the head models generated from the manually corrected segmentations (Huang et al., 2017). Similar to the results in (Huang et al., 2019), we find no statistically significant differences in the accuracy of the field predictions between the methods. This result suggests that even though there are clear differences in the electric field simulations between the methods, as shown in Table 2, these differences are not reflected in the comparison with the measurements. The large inter-subject variability in the regression fits likely explains the inconclusive result, but, as we show in the following, this variability is not a result of poor electric field simulations alone but is partly explained by the limitations of the intra-cranial measurements.

To link the intra-cranial measurements to the slope estimates from the standard regression, we plot the correlation between the strength of the recorded potential differences and the slope estimates in Fig. 8. We find a statistically significant correlation for all methods except headreco. That is, the linear relation is weak, i.e., the slope is close to zero, for the subjects where the measured signal is also weak. In contrast, the slope is steeper and closer to one for subjects where the measured signal is strongest. This implies that the slope estimates are underestimated in the standard regression analysis when there are large uncertainties present in the simulations. In fact, it is well-known that if noise in the so-called independent, or predictor, variables is unaccounted for, the regression coefficient will be underestimated (Frost and Thompson, 2000; Fuller, 1987). The problem persists even if the predicted and independent
variables are exchanged as then the noise in the measurements is ignored, or if an intercept term is added.

To exemplarily demonstrate how the regression analysis is affected by the measurement data, we look at subject P014. Subject P014 received stimulation with four different electrode positions, which allows us to untangle the effect of the measured signal at the electrodes from the impact of the head model, as the FEM mesh is fixed for all four positions.

Two of the electrode configurations (P014A and P014D, blue and black filled diamonds in Figs. 7 and 8) seem to have a good fit between simulations and measurements, based on the slope fits, for the majority of the pipelines, whereas the other two configurations (P014B and P014C, green and red filled diamonds) have poorer fits. This prompted us to conduct a full Bayesian regression analysis (Eqs. (4)–(6)), as outlined next.

4.2.2. Bayesian errors-in-variables regression
Fig. 9 shows the results of the Bayesian analysis for configurations P014C and P014D (red and black filled diamonds in Figs. 7 and 8, depicting the results of the classical regression analysis). Please see the Analyses section for a summary of the conducted analysis and

Table 3
Mean ±Standard deviation of the coefficient of determination ($r^2$), slope ($\beta$) and correlation ($\rho$) for each head modeling pipeline, across all subjects.

<table>
<thead>
<tr>
<th></th>
<th>mri2mesh</th>
<th>headreco + CAT</th>
<th>headreco</th>
<th>ROAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient of Determination ($r^2$)</td>
<td>0.506 ±0.163</td>
<td>0.580 ±0.150</td>
<td>0.567 ±0.148</td>
<td>0.554 ±0.175</td>
</tr>
<tr>
<td>Slope ($\beta$)</td>
<td>0.526</td>
<td>0.632</td>
<td>0.556</td>
<td>0.644</td>
</tr>
<tr>
<td>Correlation ($\rho$)</td>
<td>0.700 ±0.127</td>
<td>0.755 ±0.106</td>
<td>0.746 ±0.103</td>
<td>0.734 ±0.128</td>
</tr>
</tbody>
</table>

Fig. 6. The distribution of the norm of the electric field over the full brain volume for example subject P03 from ROAST and headreco + CAT. Note the differences in smoothness of the simulated fields.
Supplementary Material 1 for further details. The measured potential differences for P014C are closer to zero and look relatively noisier than for P014D. This results in a lower slope estimate and a wider 95% compatibility interval for P014C compared to P014D. This implies that some of the poor fits observed in Fig. 7 are not only due to simulation errors, but rather the fact that in the absence of clear measurement signal, the uncertainties in the simulations dominate the linear fit and cause slope estimates close to zero. This gives more insight into the inter-subject variability observed in Fig. 7 and is not revealed by the standard regression analysis. Detailed analysis, similar to Fig. 9, on all subjects can be found in Supplementary Material 2.

To quantify the differences between the slope estimates from the standard and Bayesian analysis we sampled 4000 slope estimates from the normal distribution governing the slope of the standard regression in each subject, compared those in a pairwise manner to the slope samples from the posterior distribution in each subject, and computed the probability that the standard regression slope is smaller than the corresponding Bayesian one. The probabilities computed this way are: 0.867 for mri2mesh, 0.866 for headreco + CAT, 0.869 for headreco, and 0.861 for ROAST, revealing a large probability that the slopes obtained with standard analysis will be smaller than the ones obtained with the Bayesian analysis. The distribution of these pooled differences for each method over all subjects is shown in Supplementary Material 2 (Figure S18). All distributions have medians larger than zero indicating that the slope estimates from the Bayesian analysis are generally larger than the ones obtained from the standard regression analysis. To link this to a more classical statistical analysis, we performed paired Wilcoxon signed-rank tests between the standard slope estimates (Fig. 7) and the posterior means, which resulted in p-values < 0.001 for all segmentation methods. Thus, both the Bayesian analysis and a standard pairwise test between the slope estimates indicate that the regression results in Fig. 7 and Table 3, along with the results in (Huang et al., 2019) and (Huang et al., 2017) are underestimating the true slopes as they do not account for the uncertainties involved in the simulated fields. The analysis, results, and conclusions presented in (Huang et al., 2019) and (Huang et al., 2017) would benefit from being revisited with this underestimation in mind. We note that similar analyses of the correlation coefficients (Fig. 7, left and Table 3, second row) did not show differences between the standard and Bayesian regression, indicating that the correlation estimates from the two analyses agree to a large extent on this data set. The

Fig. 8. Correlation between the 95th percentile of the absolute measured field and the slope estimates. Linear regression line without intercept shown in black. The correlation is significant for all methods except headreco and indicates that the standard slope estimates are likely biased.
Fig. 9. Bayesian regression analysis for subject P014. The first panel shows results for P014C, and the second panel for P014D. In each scatter plot the black line denotes a slope of one, the orange line is the median of the posterior distribution for the slope with the shading denoting 95% compatibility interval, and the green line is the standard slope fit (as in Fig. 7). The histograms show the posterior distributions of the slope and correlation with the median denoted as an orange solid line, the 95% interval as a dashed line, and the green solid line denoting the standard fit (as in Fig. 7). Similar plots for all subjects are included in the supplementary material.
5. Discussion

In the current work, we analyzed an openly available dataset (Huang et al., 2017) with intracranial electric potential recordings, MR scans and manually corrected segmentations to relate electric field simulations from four simulation pipelines to measured data. First, we showed that the differences in the segmentation and FEM modeling approaches in the software pipelines result in clear differences in the simulated electric field distributions. Next, we linked the simulations to the intra-cranial measurements using standard statistical analysis showing that all methods predict the measurements equally well even though the simulated fields differ up to 49%. This result was also found previously by (Huang et al., 2019). We extended this analysis using a Bayesian errors-in-variables regression showing that the slope estimates from the standard analysis are underestimated, and further demonstrating how the noisiness of the recorded data results in larger uncertainties in the regression parameter estimates.

The results highlight two important points: first, although intracranial recordings are considered the gold standard for validating computational electric field models, careful analysis of the fit between simulations and measurements is needed to avoid overly confident conclusions on how well, or poorly, the simulations match the measured data. Specifically, the noise in the measurements can affect a standard linear regression analysis leading to the incorrect conclusion that the simulations systematically overestimate the measured fields as shown in Fig. 7. Second, modeling the noise in the measurements and the uncertainty in the simulated e-field values, along with a fully Bayesian treatment of the regression analysis, is important for obtaining unbiased slope estimates and for evaluating if the variability in the slope estimates is too high for conclusive validation of the simulations. Decreasing this variability by increasing the quality of the measurements should be the main goal for future validation attempts. As exemplified by subject P014 in Fig. 9, the location of the stimulation electrodes has a large effect on how much signal is measured at the intra-cranial electrodes. The location of the stimulation electrodes should be carefully planned, as the intracranial electrodes can measure potential differences only in the plane defined by the electrode strip (Huang et al., 2017), to make conclusive validation possible. Additionally, measurements of the electric field in all three directions would likely distinguish differences between field simulations better but are difficult in practice with intra-cranial electrodes. Complementing intra-cranial recordings with new, non-invasive techniques, such as MRCDI, for volume measurements of the electric field models, careful analysis of the electric field distributions. Next, we linked the simulations to the intra-cranial recordings are considered the gold standard for validating computational electric field models, careful analysis of the fit between simulations and measurements is needed to avoid overly confident conclusions on how well, or poorly, the simulations match the measured data. Specifically, the noise in the measurements can affect a standard linear regression analysis leading to the incorrect conclusion that the simulations systematically overestimate the measured fields as shown in Fig. 7. Second, modeling the noise in the measurements and the uncertainty in the simulated e-field values, along with a fully Bayesian treatment of the regression analysis, is important for obtaining unbiased slope estimates and for evaluating if the variability in the slope estimates is too high for conclusive validation of the simulations. Decreasing this variability by increasing the quality of the measurements should be the main goal for future validation attempts. As exemplified by subject P014 in Fig. 9, the location of the stimulation electrodes has a large effect on how much signal is measured at the intra-cranial electrodes. The location of the stimulation electrodes should be carefully planned, as the intracranial electrodes can measure potential differences only in the plane defined by the electrode strip (Huang et al., 2017), to make conclusive validation possible. Additionally, measurements of the electric field in all three directions would likely distinguish differences between field simulations better but are difficult in practice with intra-cranial electrodes. Complementing intra-cranial recordings with new, non-invasive techniques, such as MRCDI, for volume measurements of the electric fields seems thus important for future validations of individualized field modeling.

Another important factor for validation is the quality of the MR scans. As shown in Figs. 2 and 3, the contrast between the tissues directly relates to the variability of the segmentations. This introduces an additional source of uncertainty on top of the measurement noise, and the two are difficult to untangle in the regression analysis. The data set analyzed here consists of a clinical population, with only T1-weighted MR scans available, where the data quality is understandably variable. Future studies would benefit from including a T2-weighted scan, which has been shown to help with skull segmentation accuracy of automated methods.
In prior articles based on this dataset (Huang et al., 2019, 2017), the comparisons between simulations and recordings have relied on a standard linear regression analysis. In this work, we reproduced the results from (Huang et al., 2019, 2017), in Fig. 7 and Table 3, where no difference between the methods was found. In (Huang et al., 2017) the authors compare simulations based on the manually corrected head models to intra-cranial recordings and, based on these, optimize the tissue conductivities on a subject-to-subject basis. In both articles, the simulations are found to overestimate the recorded potential differences based on the slope estimates from the standard linear regression. This overestimation is likely overstated due to the bias in the slope estimates and are hard to put into perspective as the uncertainties in the estimates are not assessed. This problem also links to the optimization of the tissue conductivities in (Huang et al., 2017), which relies on minimizing the squared error between the simulations and recordings, where both are assumed to be noiseless. It seems likely that this procedure provides biased conductivity estimates. Regardless of its limitations, this open-source data set is at the moment still unique and we acknowledge its importance for validating volume conductor models to intra-cranial recordings and, based on these, optimize the tissue conductivities in (Huang et al., 2017), which relies on a ring electrode versus conventional rectangular pad. Brain Stimul. https://doi.org/10.1016/j.brs.2014.02.004.


