Probabilistic maps for deep brain stimulation − Impact of methodological differences

Teresa Nordin [a, *], Dorian Vogel [a, b], Erik Österlund [c], Johannes Johansson [a], Patric Blomstedt [d], Anders Fytagoridis [c], Simone Hemm [a, b], Karin Wårdell [a]

[a] Department of Biomedical Engineering, Linköping University, Linköping, Sweden
[b] Institute for Medical Engineering and Medical Informatics, School of Life Sciences, University of Applied Sciences and Arts Northwestern Switzerland, Muttenz, Switzerland
[c] Department of Clinical Neuroscience, Neurosurgery, Karolinska Institute, Stockholm, Sweden
[d] Department of Clinical Science, Neuroscience, Umeå University, Umeå, Sweden

ARTICLE INFO

ABSTRACT

Background: Group analysis of patients with deep brain stimulation (DBS) has the potential to help understand and optimize the treatment of patients with movement disorders. Probabilistic stimulation maps (PSM) are commonly used to analyze the correlation between tissue stimulation and symptomatic effect but are applied with different methodological variations.
Objective: To compute a group-specific MRI template and PSMs for investigating the impact of PSM model parameters.
Methods: Improvement and occurrence of dizziness in 68 essential tremor patients implanted in caudal zona incerta were analyzed. The input data includes the best parameters for each electrode contact (screening), and the clinically used settings. Patient-specific electric field simulations (n = 488) were computed for all DBS settings. The electric fields were transformed to a group-specific MRI template for analysis and visualization. The different comparisons were based on PSMs representing occurrence (N-map), mean improvement (M-map), weighted mean improvement (wM-map), and voxel-wise t-statistics (p-map). These maps were used to investigate the impact from input data (clinical/screening settings), clustering methods, sampling resolution, and weighting function.
Results: Screening or clinical settings showed the largest impacts on the PSMs. The average differences of wM-maps were 12.4 and 18.2% points for the left and right sides respectively. Extracting clusters based on wM-map or p-map showed notable variation in volumes, while positioning was similar. The impact on the PSMs was small from weighting functions, except for a clear shift in the positioning of the wM-map clusters.
Conclusion: The distribution of the input data and the clustering method are most important to consider when creating PSMs for studying the relationship between anatomy and DBS outcome.

1. Introduction

Deep brain stimulation (DBS) is an established procedure for alleviating motor symptoms in diseases such as Parkinson's disease, essential tremor (ET), and dystonia. Using electrical pulses, DBS can impact the neuronal tissue and disrupt pathological signals [1]. Different targets have been suggested but the optimal locations for stimulation are still debated [2].

A common method to evaluate the spatial impact of the stimulation is to simulate the electric field around the DBS lead using the finite element method (FEM) [3,4]. The volume of tissue

activated (VTA) by the stimulation can be estimated from an electric field threshold based on single cable neuron model simulations [5]. By using a fixed isolevel it is possible to compare simulations within a patient and to combine simulations for group analysis.

The number of studies aimed to do group analysis in order to spatially evaluate the effect of DBS has increased during the last decade, commonly by creating probabilistic stimulations maps (PSM) [6]. These studies include a variety of disorders and targets where Parkinson's disease and the subthalamic nucleus (STN) are the most frequently investigated [6–11]. Most of these studies focus on finding the best anatomical area to stimulate for improving the symptom profile. Adverse effects have also been investigated, but very few of these studies include PSMs [8,12]. A typical workflow for creating PSMs starts with transforming each patient's simulations to a common space, either to an existing reference like MNI152 [13] or a group-specific template [14]. Then, voxel-wise statistics are applied where the result from electric field simulations are combined with the clinical effect. The outputs are PSMs which estimates the improvement or possibility of side effect in each voxel. Some studies also include a validation step, e.g. out-of-sample or cross-validation, to estimate the level of predictability of the created map on new patient data [6,8,15].

Numerous factors in the workflow can impact the result of a disease-specific PSM [16]. For example, many studies create PSMs based on the optimized clinical settings used in the patient's daily life, and few studies use data from monopolar review, where each contact is screened for the best outcome. Another example is the use of a binary VTA or the alternative to introduce a weighting function derived from the electric field or distance from DBS lead, as first presented by Eisenstein et al. [7]. Dembek et al. [17] used in silico data and showed a notable variation in sweet-spots size and location using five different binary methods. However, to our knowledge, there is no systematic study investigating the impact of methodological alternatives on the created PSMs using patient data.

The aim of this study was to evaluate how methodological choices impact the PSMs. A dataset of 68 ET-DBS patients implanted in caudal zona incerta (cZi) were used to investigate the influence from 1) clinical vs screening data, 2) clustering methods for extracting regions of high or low improvement, 3) sampling resolution of the electric field and 4) introduction of a weighting function. All data was analyzed and visualized on a cohort specific MRI template.

2. Materials and methods

This section is divided into three main parts. First, the patient-specific data and computations are described together with the process of moving from patient space to a common space for group analysis. Second, the methods for creating PSMs are described. Finally, the analyses for comparing methodological differences are presented. The complete processing workflow from clinical input to PSMs is described in Fig. 1.

2.1. Patient-specific data and computations

2.1.1. Clinical data
Clinical- and anatomical stimulation data for parts of this cohort has been previously published [15,18,19]. DBS contact evaluations and brain images from 77 patients were included in this study. All patients were diagnosed with ET by a movement specialist and had received cZi DBS (3387 or 3389, Medtronic, Minneapolis, MN, USA). From the 77 patients, 9 were excluded in the statistical analysis due to incomplete stimulation or improvement data. For the remaining 68 patients (31 females) the mean age at time of surgery was 63.6 years. 13 patients received bilateral DBS electrodes, 54 unilateral and one had two ipsilateral electrodes (both implanted in cZi with

different laterality). Of the DBS leads 64 were on the left and 19 on the right hemisphere. Informed consent was obtained according to the Declaration of Helsinki, and the study was approved by the Ethics Committee at Umeå University (Dnr. 2017/122−31).

The surgical procedure has been described in detail before [20]. Preoperative stereotactic T_1 and T_2 MRI images (1.5 T) were used for direct targeting of the cZi. After surgery, a stereotactic CT was performed to confirm electrode position.

Using the essential tremor rating scale (ETRS), patients were evaluated preoperatively and 1 year after surgery, both off stimulation and with the clinically used stimulation settings [21]. Each individual contact was evaluated 1 year postoperatively with standardized stimulation parameters: monopolar stimulation with 60 μs pulse width, 145 Hz, and increasing amplitude in increments of 0.2−0.3 V up to 5 V. In this manner, the stimulation was iterated for each contact to achieve the best effect on tremor and hand function based on ETRS item 5/6 and 11−14. At the same time, side effects were registered and categorized. In this paper, the side effect dizziness is used as an example. The lowest amplitude at which the highest tremor suppression occurred from the individual contact evaluations will be referred to as screening data and the clinically used settings as clinical data.

2.1.2. FEM simulation
To estimate the electrical impact of the DBS, the electric field in the tissue was computed using FEM simulation (Comsol Multiphysics v. 5.5, COMSOL AB, Sweden), based on the method from Åström and colleagues [3,5,22], see also Fig. 1b. Shortly, the volume conductor model was used with a heterogenous isotropic brain conductivity model based on the patients preoperative T_2 images and stimulation parameters. The geometry of the DBS lead was modelled and placed according to the artefact in the postoperative CT and a 250 μm peri-electrode space modelled as white matter was included [23,24]. In total, the simulations were restricted to a box of $100 \times 100 \times 100$ mm which was meshed with tetrahedral elements (mesh element size 0.02−2 mm, smallest closest to the lead). Since anodic stimulation impact the neurons to less extent, the resulting electric field from bipolar stimulation on the anodic side was scaled with a factor of 0.6 [25,26]. Based on neuron model simulations, an isolevel of 0.2 V/mm was selected to threshold the electric field since lower electric field magnitude is assumed to not impact the neurons [5,27]. Simulation results based on pulse width 90 μs were scaled with a factor of 1.33 which corresponds to an isolevel of 0.15 V/mm [5]. The scaling was required to compare simulation results spatially, since longer pulse width can stimulate neurons at lower electric field magnitude.

In total, 488 simulations were computed, 73 for clinical settings, 312 for screening settings and 103 where dizziness occurred. These results were further transformed to a common space for group analysis.

2.1.3. MRI template and electric field resampling
In order to conduct group-level analysis, spatial anatomical normalization to a template volume was performed using a methodology previously presented [14,28], Fig. 1c. It consists in the creation of an initial affine template using the MNI nLin-aSym 2009b as reference [13]. The affine template is then used as reference for iterative non-linear normalization, five times with the coarse registration settings from Ewert [29] and three times with finer registration settings from Vogel [28]. Each iteration is followed by scaling of the resulting transforms to avoid divergence of the anatomy [30]. This process outputs an MRI template, which is an average representation of all included patients, and deformation fields explaining the transform from patient space to the template. The normalization was conducted for 71 of the 77 patients. Six
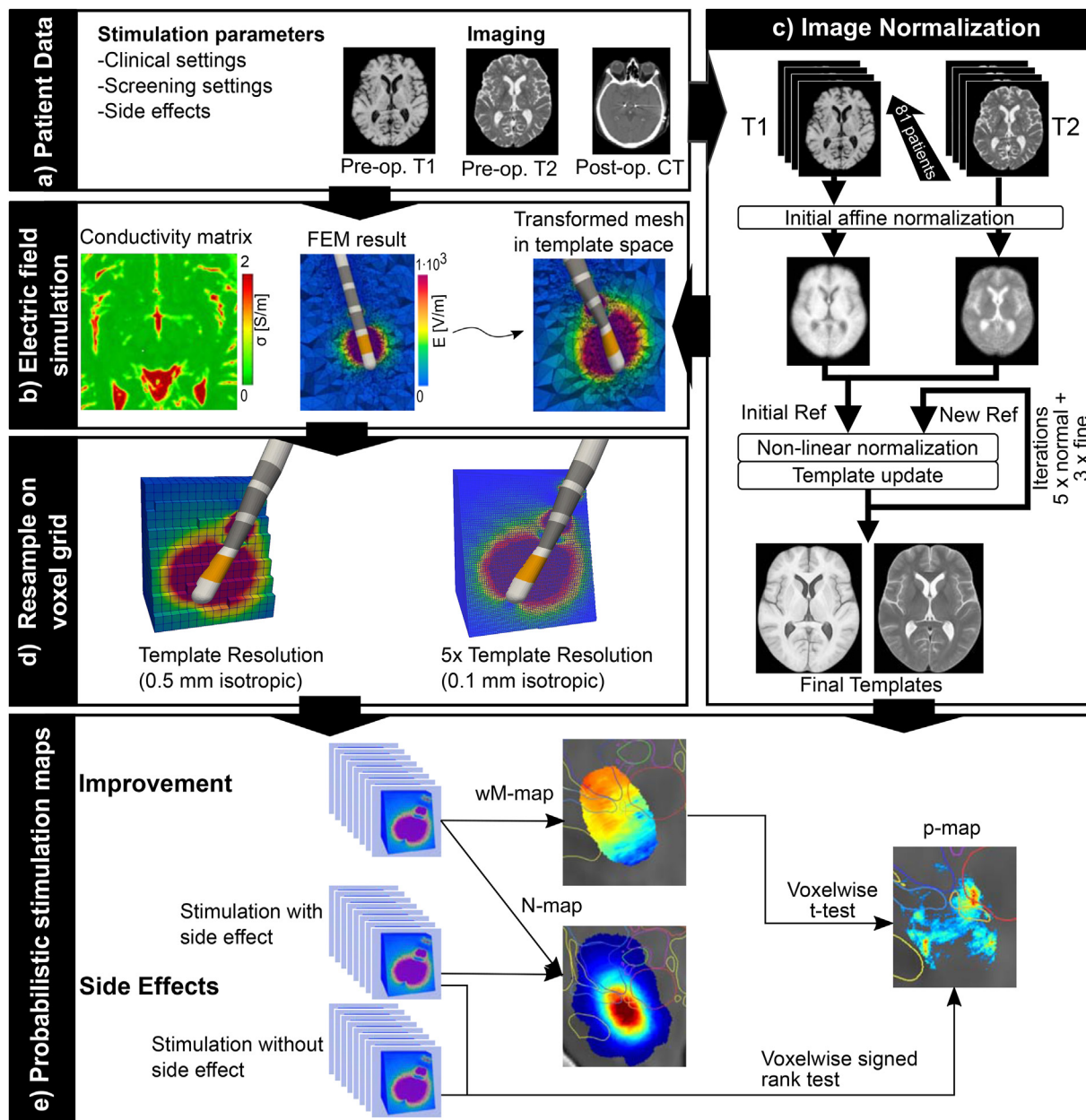
Fig. 1. Overall workflow. Workflow from patient data to electric field simulation and image normalization. The workflow starts with the a) patient data that is used for b) electric field simulation and c) image normalization. The electric fields are transformed to the normalized template and d) resampled with two different resolutions. The resampled results are combined to create e) the PSMs displaying occurrence (N-map), weighted mean (wM-map) and p-value from statistical test (p-map).

patients were excluded in the template, due to disproportionally higher resolution in the MR images compared to the rest of the group. These six patients were directly registered to the final template with the finer registration settings.

For every simulation result, the tetrahedral meshes were transformed to the template space by applying each patient's deformation field produced during the normalization process, Fig. 1d. The transformed electric fields were then resampled to voxel grids in the template space with Gaussian distance weighting based on the 4 mesh points closest to the voxel position on the grid. Since simulations were not computed within the lead, corresponding voxel values were interpolated to avoid holes in the results. To reduce data size, the resampling for each simulation result was limited to the region with electric field magnitude above 0.16V/mm (0.2V/mm x 0.8 margin factor). Two different sampling

resolutions were used, one identical to the anatomical template (500 μm isotropic) and one five times higher (100 μm isotropic). In order to provide the outlines of anatomical structures of the deep brain in the visualizations, the deep brain atlas from Morel [31] in its digital version [32] was registered to the group template using ANTS-SyN with the finer registration settings.

The normalization, transforms and resamplings were done on the Sigma Cluster System from the National Supercomputer Centre at Linköping University.

2.2. Probabilistic stimulation maps

Electric field simulations from all patients were combined to create PSMs which visualizes voxel-wise descriptive statistics and statistical tests. The descriptive statistics were mean value (M-map),

weighted mean value (wM-map) and occurrence (N-map). The statistical tests resulted in p-value maps (p-map), Fig. 1e. The maps were computed for the screening data, clinical data and where dizziness occurred for the left and right implantation sides separately. To increase the statistical power, voxels with less than 10% of the data or 5 simulations, whichever greatest, were removed from the maps except for the N-map. All statistical computations were done in MATLAB (v. 2021a, The MathWorks Inc., Massachusetts, USA).

2.2.1. Improvement

In each voxel, the mean improvement in tremor reduction (*TR*) weighted with the electric field was computed, resulting in a wM-map. To promote a lower amplitude, i.e. smaller VTA, the electric fields were divided by the stimulation amplitude which effectively create a distance weighting according to

$$\widehat{TR}_i = \frac{\sum_k \left(\frac{E_{ik}}{V_k} \cdot TR_k \right)}{\sum_k \frac{E_{ik}}{V_k}} \qquad (1)$$

where E (V/mm) is the electric field magnitude, V (V) is the stimulation amplitude, TR (%) is the tremor reduction, i denotes the voxel and k the simulation. The baseline for the TR was the 1-year postoperative OFF stimulation score.

A double-sided test was constructed to find voxels representing high or low clinical improvement by using the following *t*-test statistics

$$t_i = \frac{\widehat{TR}_i - 70}{\sqrt{\frac{s_i^2}{N_i}}} \qquad (2)$$

where N is the number of simulations impacting voxel i and s is the standard deviation. The test is designed to find voxels significantly deviating from 70% improvement, which is the mean of all stimulations from the screening sessions. The standard deviation of the voxel-wise mean improvement was calculated as a weighted standard deviation.

$$s_i = \sqrt{ \frac{ \sum_k \frac{E_{ik}}{V_k} \cdot \left(TR_k - \widehat{TR}_i \right)^2 }{ (N_i - 1) \sum_k \left(\frac{E_{ik}}{V_k} / N_i \right) } } \qquad (3)$$

This test resulted in a p-map, where all voxels have a corresponding p-value from the statistics where p < 0.05 was regarded as significant. To elucidate if a voxel was correlated to improvement higher or lower than 70%, the p-values were multiplied with the sign of the t-statistics creating a signed p-map. For comparability, same threshold was used for both the screening and clinical datasets.

2.2.2. Side effect

The side effect, in this case dizziness, was tested in each voxel by a double-sided Wilcoxon signed rank test. The rank was given by comparing the electric field magnitude from stimulations with or without occurrence of dizziness. The simulations without dizziness were created by reducing the amplitude of each stimulation resulting in dizziness, assuming lower amplitude will not generate the side effect. The amplitude reduction was 1V but for side effects occurring at <1V the non-side-effect voltage was chosen to 0.1V. All voxels with p < 0.05 were considered as significant.

2.2.3. Permutation test

To correct for type 1 error of multiple comparisons, due to a multitude of voxels tested on the same data, a permutation test was implemented where the effect values were permuted 200 times, as

first presented in Ref. [7]. For the screening data, the effects were first permuted between the DBS leads and secondly within the leads [8]. For side effect and clinical settings, all effect values were permuted in one step. For each permutation, the statistical test (Eqs. (1)–(3)) was re-performed. To test if a random permutation gives higher significance than the true p-map, an overall statistic (*Q*) was created and calculated for each permutation

$$Q = \sum_i \{ -log_{10}(p_i) \text{ if } p_i < 0.05, \ 0 \text{ if } p_i \geq 0.05 \} \qquad (4)$$

where p_i is the p-value in voxel i [7]. The Q statistics were used to calculate an overall p-value according to

$$p_{tot} = \frac{\sum_n \{ 1 \text{ if } Q_n > Q_{true}, \ else \ 0 \}}{N_{perm}} \qquad (5)$$

where N_{perm} is the number of permutations, Q_{true} is the *Q*-value for the true p-map and Q_n the *Q*-value for the n:th permutation. A $p_{tot} <$ 0.05 assumes an overall significance of the true p-map.

2.2.4. Visualization

For clinical, screening and side effect settings the different PSMs were created. The N-maps were created as a voxel-wise sum of all simulations with electric field above the 0.2 V/mm isolevel and used to compare stimulation location. The p-map was thresholded at p < 0.05 to create clusters of significantly higher or lower improvement, clusters separated based on the sign of the t-statistics. In the visualizations, the natural logarithm of the inverse p-map was used to improve the dynamic. For dizziness, only p-maps and N-maps were created. All maps were displayed on the MRI template, section 2.1.3, together with labels and outlines from Morel's Atlas [31].

To evaluate the stimulation maps in correlation to the anatomy, the overlaps of each significance cluster with the structures in Morel's Atlas were computed.

2.3. Comparison of methodological differences

Four different parameters were altered to evaluate the impact on the PSMs. In general, visual inspection and general comparison of outputs of the different maps were conducted in all analyses.

First, the dependency on input data, i.e. screening vs clinical data, was evaluated. The PSMs were created based on the different data sets and the voxel-wise difference in the wM-maps from screening and clinical data were computed. Only the voxels included in the analysis for both settings were used. The voxel differences were used to create 99% confidence intervals where a zero difference implies no impact of the input data.

Second, different methods for creating high or low improvement clusters were compared. The method based on the p-map was compared to clustering based on the 10% voxels with highest and lowest value in the wM-map. The intersecting volumes between the p-map clusters and clusters based on the percentile method were computed for quantification.

Third, the impact of sampling resolution was compared by first upsampling the result from template resolution to the higher resolution. The differences in the wM-maps and p-maps between the high resolution and template resolution results was computed in each voxel. These differences were also used to compute 99% confidence intervals.

Fourth, the PSMs based on weighting function were compared to PSMs from binary VTAs. Based on the presented clustering method, intersecting volumes of the binary and weighted clusters were calculated. Again, the voxel-wise differences between the wM-maps and M-maps, and the different p-maps were calculated

and used for 99% confidence intervals. For these comparisons all PSMs were re-computed but instead of the electric field, all voxels above 0.2 V/mm were set to 1.

All confidence intervals were based on double sided Student t-test.

3. Results

The generated MRI template with labels from Morel's atlas [31] is found in Fig. 2. All PSMs and volumes are presented at 5 times template resolution if not otherwise stated. For evaluation of improvement, only the left side p-map from the screening data survived correction for multiple comparisons (Table 1) and therefore the visualizations and comparisons are primarily presented for the left side. An overview of the data included can be found in Appendix A, Table A1 and Figure A1, and visualization for all improvement PSMs, in Figure A2.

3.1. Impact of methodological differences

3.1.1. Input data

The different PSMs, N-map, wM-map, and p-map, for the clinical and screening data on the left side are presented in Fig. 3a. Clinical settings provide a more focal N-map and higher voxel-wise improvement values compared to screening settings which is an expected result from the optimization of clinical settings. The wM-maps and p-maps differ for the clinical and screening data both slightly in the position of the highest improvement location and the level of improvement.

The voxel-wise mean difference in improvement between the input datasets was 12.4 and 18.2% points for left and right side respectively (Fig. 3b and c), which agrees with the skewness in the visualizations.

Investigating the p-map, the clinical data has a large cluster (312 mm^3) of voxels significantly above 70% improvement, while the low improvement cluster is very small (0.3 mm^3), Fig. 4b and c. This skewness was expected and is a reinforcement of the result in the wM-map (Fig. 3a). The overlaps between p-map clusters and different anatomical structures (Figure A2 c-d) show that both screening and clinical data on the left side have high coverage of the high improvement cluster with VLpv and VM. The main difference is that the clinical data high improvement cluster overlaps 11% of the STN volume, which is a much larger overlap than for the screening data (1.5%).
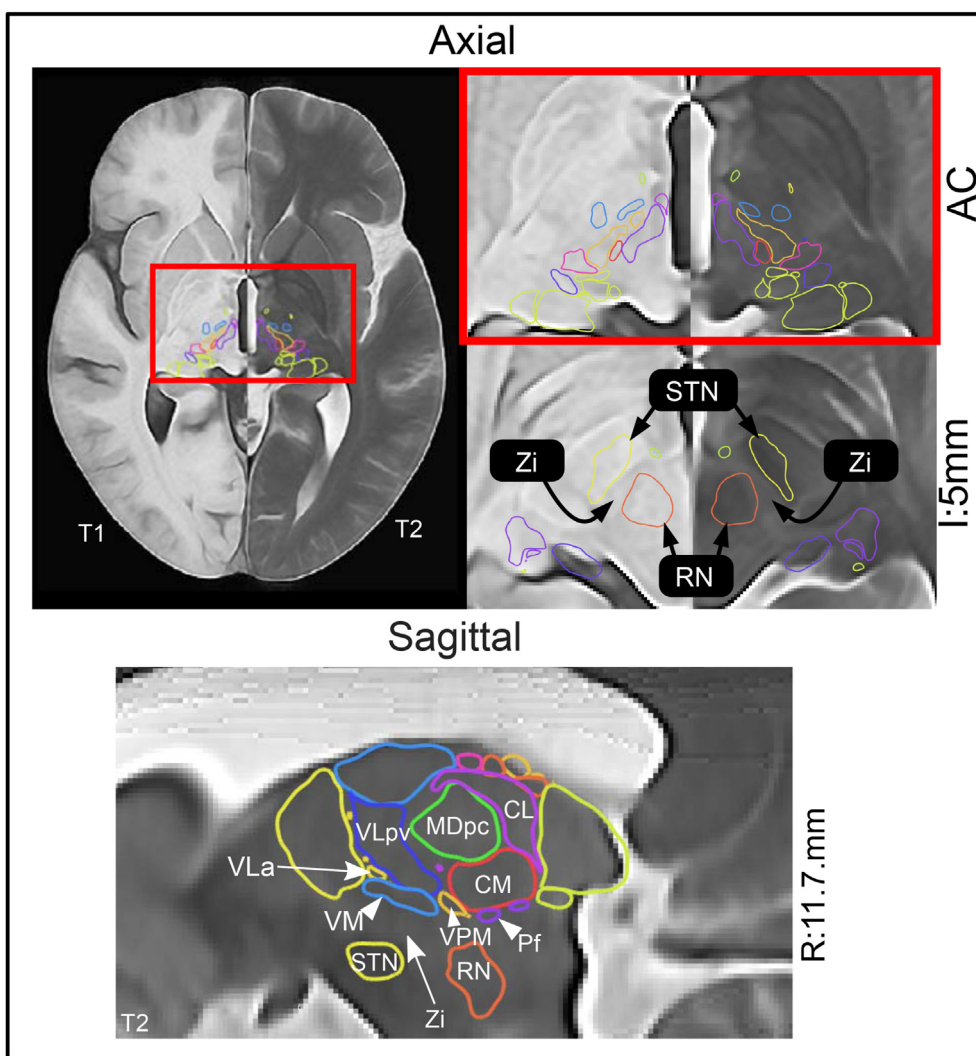

Fig. 2. MRI template. Group-specific anatomical template with labels from Morel's atlas overlayed. Top: Axial slices with left and right side displayed in T$_1$ and T$_2$ MRI respectively. Bottom: Sagittal slice in T$_2$. Note: Vim is located in the inferior part of VLpv and Zi is indicated on the MRI even though it is not delineated.

Table 1
p-values from the permutation tests. L = left, R = right, * denotes significant result.

Mode	Side	Template resolution	5x Template resolution	Template resolution	5x Template resolution
		Weighted		Binary	
Screening	L	0.005*	0*	0*	0*
	R	0.18	0.185	0.14	0.155
Clinical setting	L	0.815	0.785	0.665	0.69
	R	0.7	0.63	0.76	0.75
Dizziness	L	0*	0*	0*	0*
	R	0.02*	0.055	0.045*	0.045*
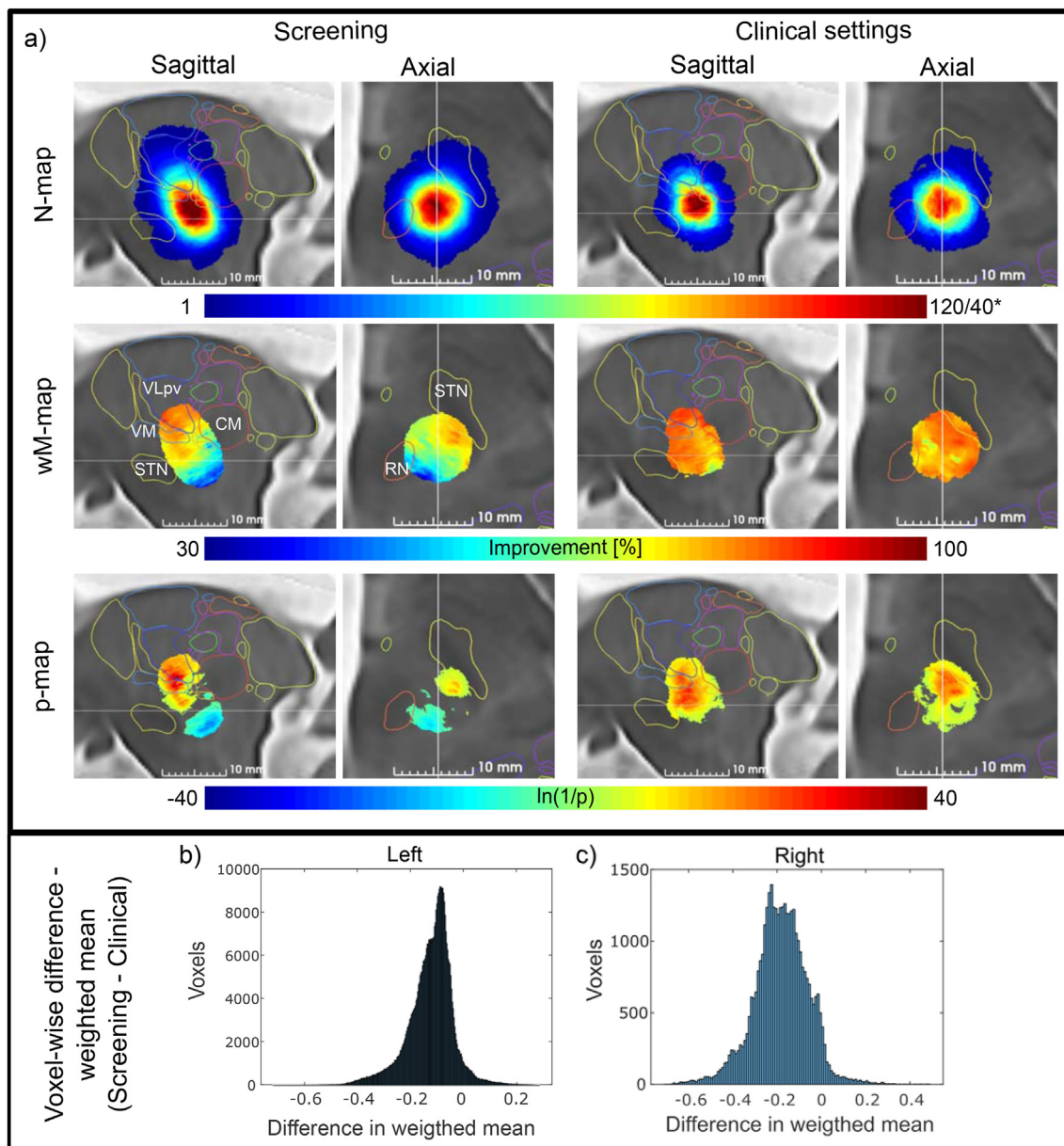
Fig. 3. Impact of input data. The impact of different input data is shown a) visually for the left side with the N-map, wM-map and p-map on the T$_2$-MRI template. The top two left columns show the screening data in sagittal and axial views and the two right columns show corresponding views for the clinical data. The N-maps spans from the inferior border of the STN to the mid thalamus with some superior outliers. The wM-maps and p-maps present different trends for the different input data. Highest improvement based on the screening data is around the lower border of thalamus while the highest improvement for the clinical settings is in the subthalamic region. As seen in the wM-maps, the clinical data is skewed towards higher improvement, i.e. more voxels shifted towards red. For both clinical and screening, the wM-maps have lower improvement close to RN, seen as a shift towards blue. Histograms of the voxel-wise difference in weighted mean (negative values means higher value in the clinical settings) are shown b) for the left side and c) for the right side. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
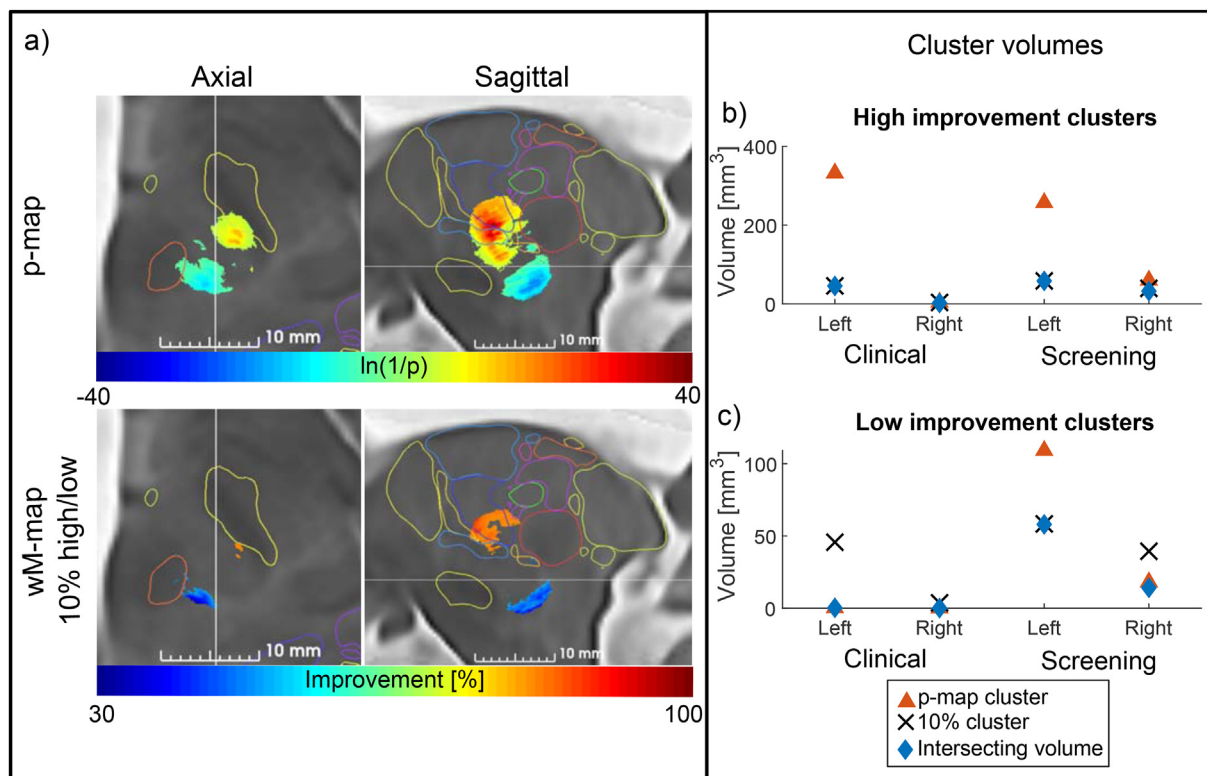
Fig. 4. Clustering methods. a) Visualization of the clustering from left side screening based on p-map and the 10% highest or lowest improvement voxels in the wM-map. The total volumes for all clusters together with the intersecting volume based on the two methods for b) the high improvement clusters and c) low improvement clusters. All results are from the maps at 5 times template resolution.

3.1.2. Clustering methods

Fig. 4 presents the high and low improvement clusters based on the significance test and the wM-map clusters based on the voxels with 10% highest or lowest improvement. The p-maps consistently generated larger clusters of high improvement (range: 4–312 mm^3) compared to the wM-map clusters (range: 2–58 mm^3), Fig. 4b. For low improvement volumes, the wM-map clusters (range: 2–58 mm^3) were larger than the p-map clusters (range: 0–109 mm^3), except for left side screening, Fig. 4c. In 5 of the 8 cluster comparisons, the intersecting volume was >99% of the smaller cluster volume. The differences between the methods can be described by the difference in cluster volume and that the voxels of higher significance are more centralized than the percentile clusters.

3.1.3. Resolution

Fig. 5 shows the results of sampling the electric field at different resolutions. The higher resolution generates a smoother result, but visually the same general trends can be seen (Fig. 5a) and the resulting p-map cluster volumes are very similar (Table 2). However, the difference in mean and p-value for each voxel generated a distribution slightly shifted from zero for 7 of 10 analyses (Fig. 5b–d). The 99% confidence intervals of the voxel-wise difference in the p-maps are shown in Table 2 and in the wM-maps in Table 3. The confidence intervals showed significant difference except for p-maps for dizziness (left), screening (right) and clinical setting (right).

3.1.4. Weighting

The impact of using the electric field norm as weighting function compared to binary VTA is visualized in Fig. 6. Even though the voxel values in the binary M-map are shifted towards higher improvement (Table 3) the weighted statistical tests reach lower p-values (Table 2) and larger volumes of significance in 8 of 10 analysis (Fig. 6b and c). The impact on the p-maps was similar for both improvement and dizziness. The largest differences were when creating clusters based on the 10% most extreme voxels where the intersections between binary and weighted clusters were 37–59% of the binary VTA for high improvement and 29–81% for low improvement.

3.2. Dizziness

Fig. 7 presents the N-map and p-map for dizziness and for comparison the N-map from screening of improvement. The dizziness N-maps are similarly located as the screening of improvement but with less spread (Fig. 7a). For dizziness, the significance test survived multiple comparison for both sides at template resolution but only on the left side at high resolution (Table 1). The p-map (Fig. 7c) forms a shell-like structure that spans from STN up in thalamus. The volumetric overlaps between p-map cluster and different anatomical structures (Fig. 7d), show overlap with several structures in the thalamic and subthalamic area. VPM, VM and VLpv are most covered by the dizziness cluster, but VPM is the only structure covered to more than 10%. CM has a very high overlap on the left side, while almost none on the right side and the opposite is seen for STN.

4. Discussion

In this study we present a method for creating and comparing PSMs for improvement and side effects in a cohort of ET patients. All results were visualized on a cohort specific MRI template. Four different PSMs (N-map, M-map, wM-map, and p-map) were used
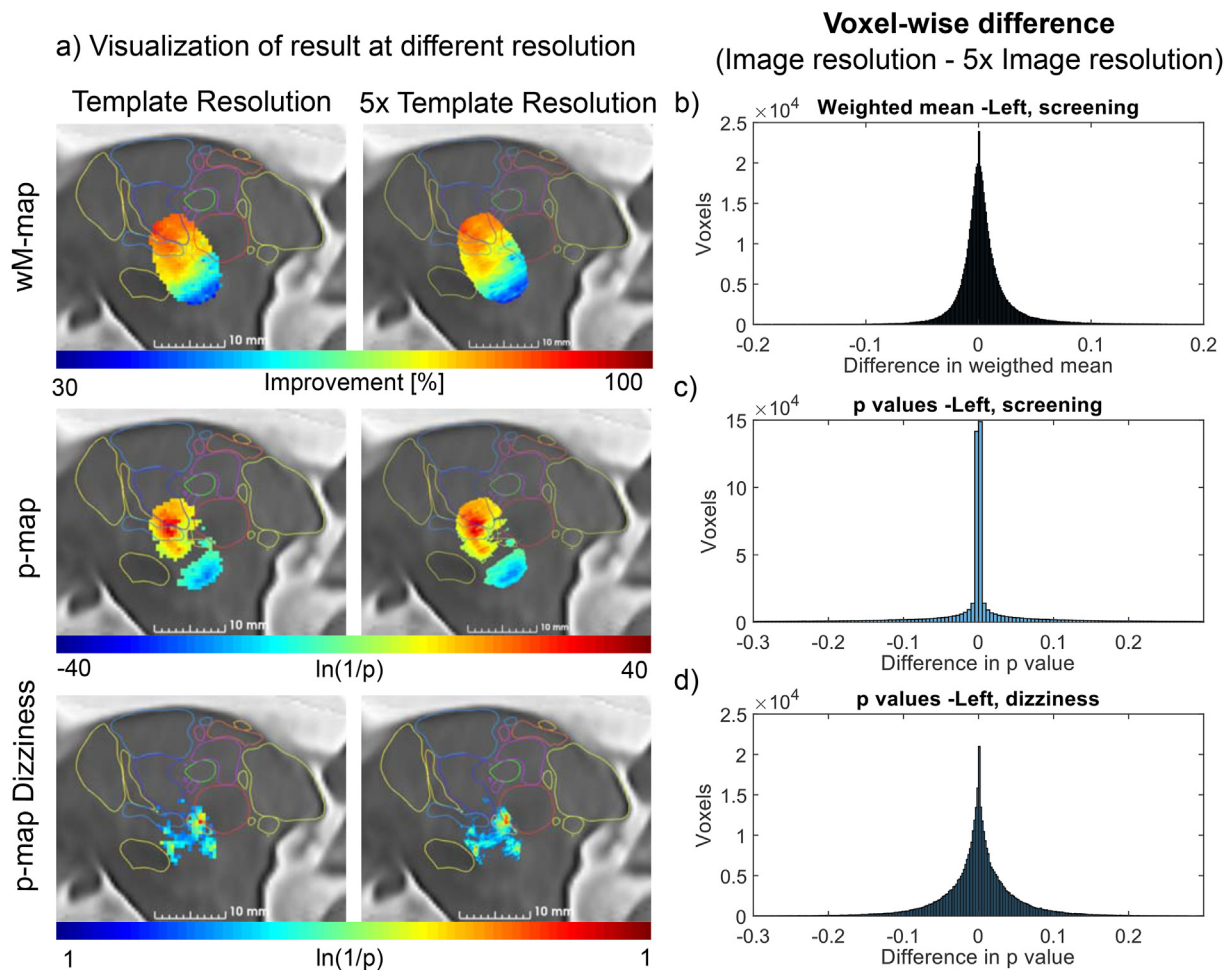
Fig. 5. Impact of resolution. a) Visualization of difference in resolution on the T_2-MR template using left side screening in the two top rows, first wM-map and second p-map, and p-map for left side dizziness in the bottom row. First column displays the image resolution and second 5 times image resolution. The voxel-wise difference is shown for the left side screening in b) wM-map and c) p-map and for dizziness d) p-map.

to evaluate the impact of using different input data, clustering methods, sampling resolution of the electric field and the use of weighting function. The result shows that all these factors impact the maps and the implications of them, but to a varying extent as summarized in Fig. 8.

4.1. Impact of input data

The improvement was evaluated both using clinical stimulation settings and settings from screening sessions where a large difference can be seen in the wM-maps and p-maps, Fig. 3. Since the clinical data are optimized for improving the symptom, it is natural

that the wM-map shows a shift towards higher improvement. However, only the screening data for the left side survived correction for multiple comparisons. The permutation strategy is useful for limiting type I errors but gives a global significance and cannot give implication on local result. Another result from the permutation strategy is that highly homogeneous data will not reach significance, which was the case for the clinical data in this study (Table 1). Since the clinical data is more biased towards higher improvement, the possibility of receiving high voxel mean value by chance increases. This also emphasize one problem with using clinical data for creation of p-maps especially for disorders such as ET where the improvement is very homogenous.

Table 2
Volumes of significance from statistical test from sampling at template resolution and 5 times template resolution. Also, confidence intervals for voxel-wise difference in p-value for the different resolutions and the difference in weighted and binary p-maps at 5 times template resolution. L = left, R = right, *Denotes significant difference between tested maps.

Mode	Side	Significance volume high/low improvement [mm³]		Confidence intervals - Voxel-wise difference p-value	
		Template resolution	5x Template resolution	High vs Low resolution	Weighted vs Binary
Screening	L	256/109	257/109	$-0.002; -6.84 \cdot 10^{-4}$*	$-0.015; -0.014$*
	R	62/18	60/18	$-0.001; 3.34 \cdot 10^{-4}$	$-0.090; -0.088$*
Clinical setting	L	330/0	332/0.2	$2.0 \cdot 10^{-4}; 0.001$ *	$0.023; 0.024$*
	R	4/0	4/0.01	$-2.02 \cdot 10^{-4}; 0.005$	$-0.30; -0.29$*
Dizziness	L	176/0	174/0	$-1.84 \cdot 10^{-4}; 2.36 \cdot 10^{-4}$	$-0.016; -0.016$*
	R	40/0	39/0	$1.95 \cdot 10^{-4}; 0.001$*	$-0.035; -0.034$*

Table 3
Confidence intervals for voxel-wise difference in weighted mean or comparison of template resolution sampling with sampling at 5x template resolution. L = left, R = right, pp = percentage points, * Denotes significant difference between tested maps.

| Mode	Side	Confidence intervals − Voxel-wise difference, weighted mean [pp]	
		High vs Low resolution	Weighted vs Binary
Screening	L	0.47; 0.49*	−0.45; −0.42*
	R	0.74; 0.77*	−0.02; 0.02
Clinical setting	L	−5.26; −5.16*	−0.52; −0.49*
	R	5.81; 6.07*	5.55; 5.75*

4.2. Impact of clustering methods

Several groups have investigated the possibility to create improvement maps in DBS. Some conducted the statistical test against a null hypothesis of 0 improvement [7,12,33], while others suggest more conservative approaches since some improvement is usually expected from all DBS contacts [8]. In this study, we choose to test against the mean improvement found during the screening sessions of the cohort. The intention of using the mean

improvement was to single out the voxels that represented extreme responders. The selected value to test against is an important aspect when interpreting the data, since the voxels that did not meet significance do not imply bad improvement, but rather that the improvement is not significantly different from 70%, which was the chosen threshold in the statistical test.

Regarding side effects, the implemented method is intended to find the border where side effects start to occur. While it would be optimal to be able to retrieve a region as for the improvement, this is not possible with this type of data since the stimulation settings are only increased until the side effect appears rather than until they are maximized as for the desired clinical effect.

Two common approaches used to find an optimal location for stimulation are either a voxel-wise statistical test [8,34] or extracting the voxels with mean improvement above the 90th percentile [12,35]. As presented in Fig. 4, the similarities are clear since the intersecting volume almost totally covers the smallest generated cluster. However, the cluster volumes retrieved differ to a large extent within and especially between the methods. The percentile-based clusters will always generate a descriptive view of the extreme voxels in the dataset, however with an unknown level of confidence and no consideration to the data variance. Due to the
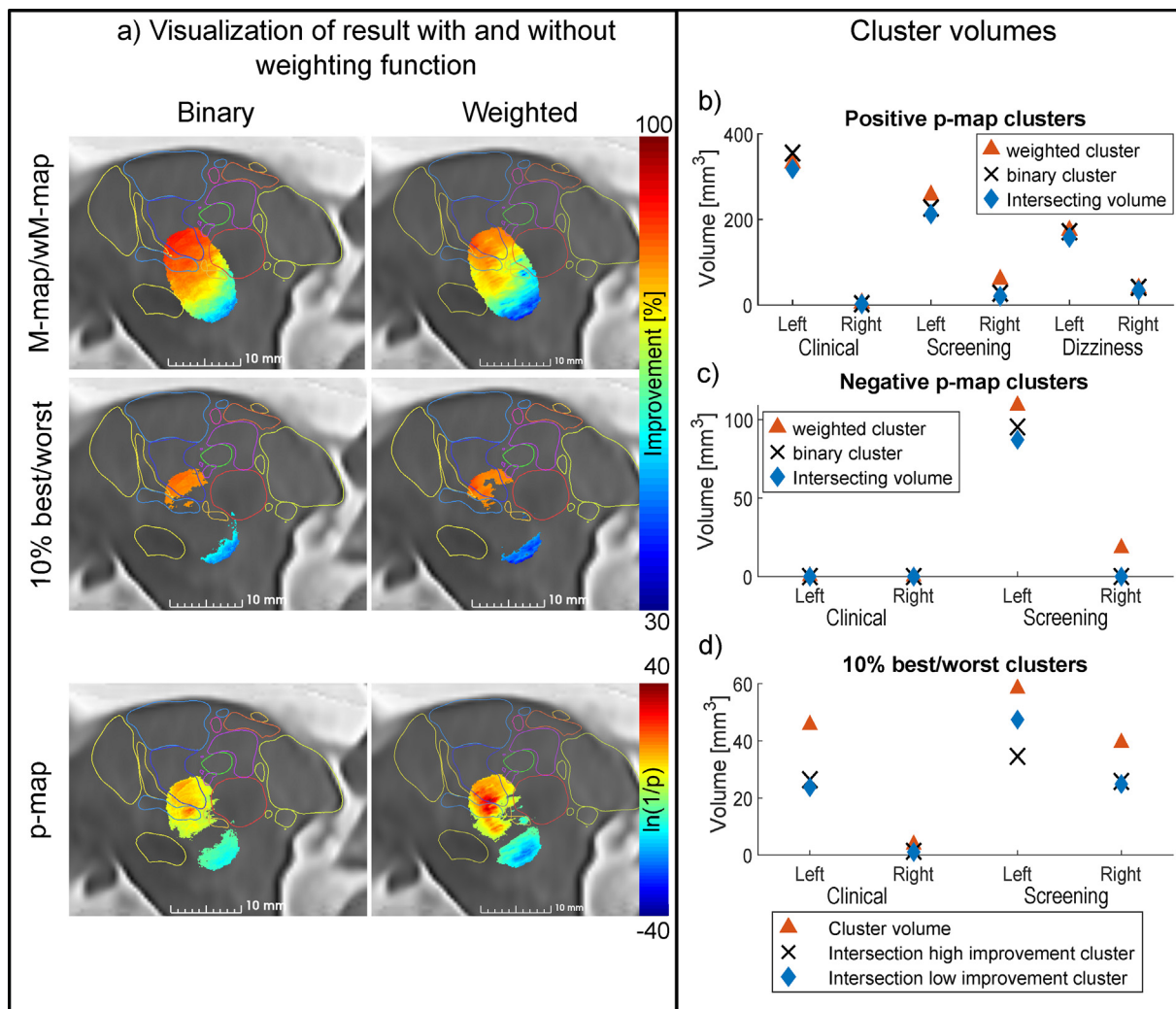
Fig. 6. Impact of weighting function. Visualization of the difference in using binary volume of activation compared to weighting with the electric field. a) Shows the M-map/wM-map in first row, the corresponding 10% best/worst cluster in the second row and p-map on last row. All visualizations are for the left side screening overlayed on the T2-MR template. The p-map cluster volumes together with intersection volume from the weighted vs binary p-maps are shown in b) for high improvement cluster and c) for low improvement clusters. d) Shows the volume for the 10% extreme voxels from the wM-map/M-map together with the intersection volumes for binary vs weighted. Note, the 10% volumes are the same size for both weighted and binary since they are only based on the number of voxels where at least 5 stimulations have an electric field above 0.2 V/mm.
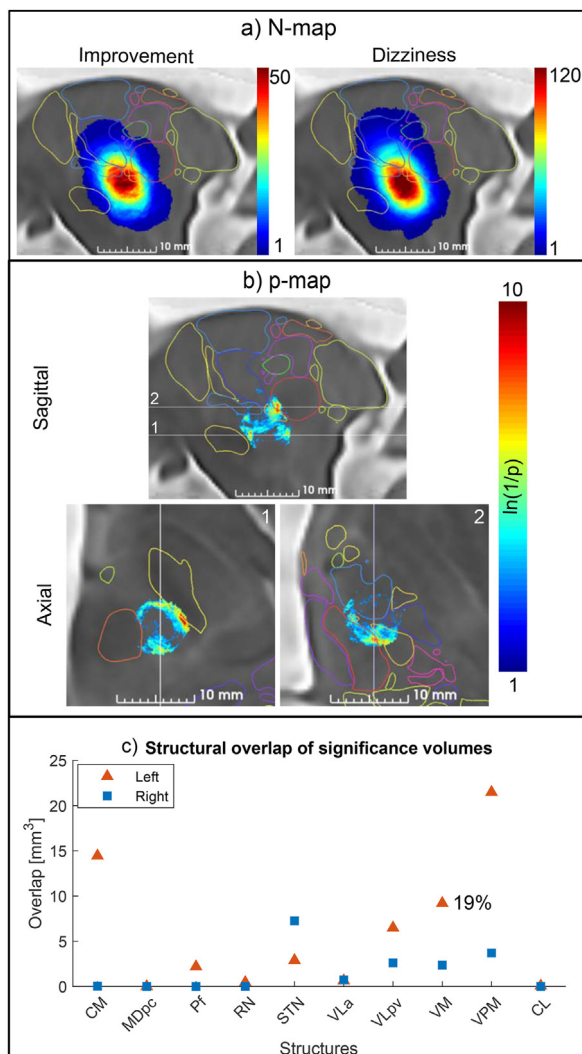
Fig. 7. Dizziness. Overall result for dizziness computed at 5 times image resolution. a) Visualizations in sagittal view from the dizziness N-map overlayed on the T_2-MR template compared to the screening data for improvement. b) Sagittal view and two axial slices of the p-map showing significance area spanning from STN up to high in Thalamus. c) Shows the overlap of the p-map with different structures in Thalamus and subthalamic area for both sides together with the % overlap of each structure where overlap >10%. Note, side effect only shows a positive probability of occurrence.

Fig. 8. Overall result of the different comparisons (input data, clustering method, weighing function and sampling resolution) ordered according to the magnitude of impact of the resulting PSM.

methods dependency on the total analyzed volume, it is also sensitive to the spatial distribution of the input.

Dembek et al. [17] used in silico data to compare 5 methods of creating sweet-spot clusters. While they found a notable variation between the methods, the screening data cluster sizes for methods using voxel-wise statistical test were generally larger than the clinical data clusters. This is in contradiction to the result found in this study, Fig. 4. One important difference is that our study uses patient data where the clinical setting is not the same as the best setting from the screening. Also, for comparability, our study used a fixed 70% null-hypothesis instead of adapting it to the dataset and therefore a larger volume for the clinical settings can be expected. However, our study together with Dembek et al. [17] highlights important aspects related to the validity of probabilistic mapping.

4.3. Impact of resolution

The resulting maps only showed small differences depending on the resolution where same overall trends were visible and the extracted volumes were similar (Fig. 4, Table 2). However, the higher resolution made details easier to see and created a smoother result without having to apply any filtering. Also, the voxel-wise difference between the maps with template resolution and 5 times template resolution showed a significant difference in 7 out of 10 cases, both for wM-maps and p-maps (Tables 2 and 3). Some differences were expected since the higher resolution can better capture the dynamic of the electric field. It is questionable though if these differences have any clinical importance due to the low magnitude, where highest confidence interval bound for weighted mean was 6% points and for p-value 0.005. What should be noted is that the resampling in this case was done after transforming the electric field to the template space to minimize errors. If the resampling had been done before the transform, which is the most common solution, the impact would have increased. This increasing impact is due to the resampling being a non-linear operation and the transform would have a larger averaging effect at lower resolution.

The main drawback with using high resolution of the electric field is the large increase in computational cost and hardware requirements. However, if that is not of importance for a project, the high resolution is probably a better choice.

4.4. Impact of weighting

The electric field was introduced as a weighting function when creating the PSMs, because higher electric field magnitude is expected to impact neurons of smaller size [5]. In the brain, most neurons are distributed around 1 μm in diameter. Axons up to 3−4 μm, which are assumed to be stimulated at the threshold 0.2 V/mm, are more rare [36,37]. However, combining stimulation with anatomy and clinical response have shown indications that the neurons are starting to be impacted by the stimulation at 0.2 V/mm [27,38]. One problem with just using the electric field as a weighting function is that higher amplitude settings will give much higher electric field values close to the lead. Giving a very high weight to the voxels closest to the lead is not reasonable from a clinical point of view since the higher amplitude is used to reach further from the lead. Moreover, lower amplitudes are generally preferred for energy efficiency and widening of the therapeutic window. Therefore, a punishment to high amplitude was added by dividing the electric field with the voltage setting, similar to what has been done by others [6,12].

Few studies have used weighting functions in PSM creation or prediction algorithms. Eisenstein et al. [7] presented a PSM method with a Gaussian weighting function but did no comparisons with

binary VTAs. Horn et al. [39] illustrated the difference in correlation between improvement and overlap of the STN for PD patients with and without a weighting function based on the electric field. The result indicated a stronger correlation when adding the weighting function, but the effect was not tested for significance. Our study indicates that lower voxel-wise p-values and larger volumes of significance can generally be found when introducing weighting. The largest difference was though for the clustering using the percentile method. Even if the total volumes were the same, the intersecting volumes were in the worst case only 29% of the binary cluster volume. This difference in cluster location implies that the percentile-based clusters are less stable against approximation with a binary volume of activation compared to p-map clusters.

4.5. Finding location for improvement and side effects

Clinical analysis of tremor reduction for 50 patients of this dataset is published in an earlier paper [18]. Similar to that study, the main trend was less improvement for the most distal contact. Higher improvement could be found in a very large volume including structures like STN, VLpv, VM and VPM, though with a trend on increasing voltage with more superior contacts, Figure A1. A recent review showed a notable variation in the optimal stimulation spots [40], suggesting that good improvement can be found in many places in the inferior part of the VLpv (the region where Vim is located) and the posterior subthalamic area (PSA). In the present study, the region with most improvement differs between screening and clinical setting. The variation is due to the input data since the mid contacts were chosen more often in the clinical setting (Figure A2b), together with the possibility to retrieve good improvement in a large volume of the region.

When improvement is easily found, the absence of side effects will be of higher importance. Dembek et al. [12] also investigated the occurrence of dizziness in patient with Vim-Zi DBS and found the significance region in anterior Zi, the VPM and PM, using nomenclature from Mai atlas [41]. Similarly, our study found the VPM to be largely covered and Zi partly covered (Zi is not delineated in the atlas used in this study) (Fig. 2) but also a large coverage of VM and VLpv. The structures stimulated when dizziness and improvement occur are very similar, but the improvement clusters are more focal and the dizziness clusters are present at the outer region. Since many structures are impacted, a clear structural correlation for dizziness cannot be established, which can relate to the side effect being diffuse and probably of multifactorial origin.

4.6. Limitation and future aspects

The main limitation for the PSM creation is the strong skewness towards higher improvement in the clinical data. This limitation is also present but to a lower extent for the screening data since only the best improvement for each contact had been collected. Although the dataset is quite large from a DBS perspective, many of the analyses suffer from limited amount of data, especially for the right side. One option could have been to transfer all simulations to the same side, but this was not implemented to not make any assumptions of laterality independence. One possibility to overcome the bias limitation would be to have improvement evaluation for multiple amplitudes per contact [12]. However, full test on each amplitude would extensively increase the time and therefore a more automatic procedure, like accelerometer measurements, would be preferable [42]. Regardless of the data limitations, the comparisons between methods are still valid and the dataset is similar to other publications.

In the future, more studies should be focusing on predictability. There are several studies on this topic for DBS applied on different disorders [8,15,35,43]. Many find significant correlations, but the models usually describe the interpatient variance poorly. A promising addition is to include structural or functional connectivity, but more research is required to find solid sets of predictors for the different disorders treated with DBS.

5. Conclusions

In this study we have investigated how different parameters impact the result when creating different probabilistic stimulation maps and searching for the optimal or less optimal locations for stimulation. We found that the electric field resampling resolution alter the PSMs to some extent, and higher resolution should be used if possible. Slightly higher impact on the maps was seen by introducing a weighting function and should especially be used for analysis on mean improvement visualizations. The largest impact was observed for the choice of the input data and clustering method. The observed impact of input data also implies that the use of multiple stimulations per position may further alter the result. Therefore, when selecting these parameters, care should be taken into the variance and possible bias in the data and the level of confidence of the resulting clusters.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

CRediT authorship contribution statement

Teresa Nordin: Methodology, Software, Formal analysis, Visualization, Writing − original draft. **Dorian Vogel:** Methodology, Software, Visualization, Writing − review & editing. **Erik Österlund:** Resources, Data curation, Writing − review & editing. **Johannes Johansson:** Supervision, Writing − review & editing. **Patric Blomstedt:** Resources, Writing − review & editing. **Anders Fytagoridis:** Conceptualization, Resources, Data curation, Writing − review & editing. **Simone Hemm:** Conceptualization, Supervision, Writing − review & editing. **Karin Wårdell:** Conceptualization, Supervision, Resources, Funding acquisition, Writing − review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Patric Blomstedt is consultant for Abbott, Boston Scientific, Koh Young and shareholder in Mithridaticum AB.

Karin Wårdell is shareholder in FluoLink AB.

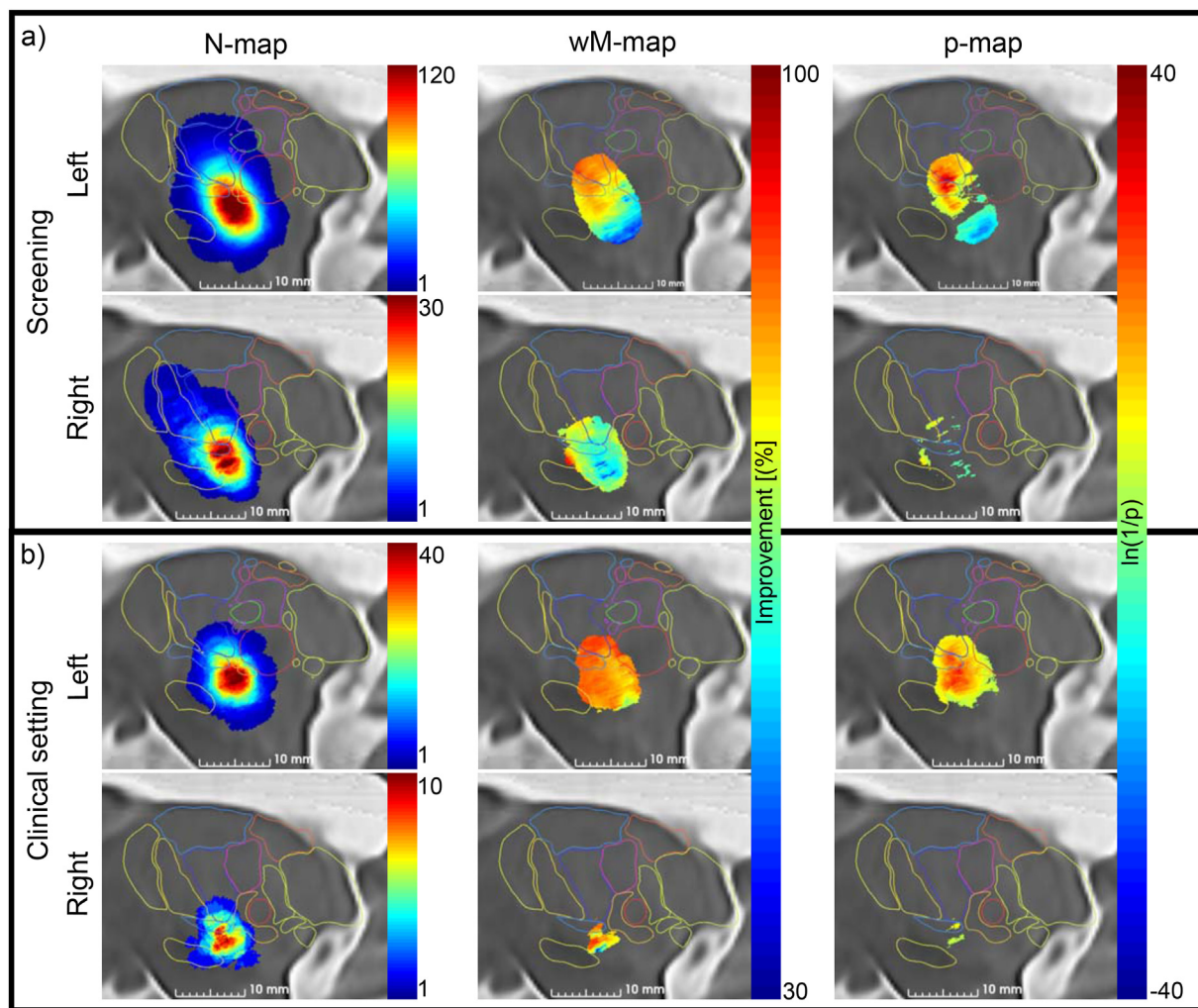## Appendix A. Patient stimulation data

Table A1
Overview of patient data. Values are noted as mean ± standard deviation. The improvement is based on ETRS score item 5/6 and 11−14. L = left, R = right.

Mode	Side	N contacts (pat)	Amplitude [V]	Pulse width [µs]	Frequency [Hz]	Improvement [%]	Stimulated volume [mm³]
Screening	L	248(63)	2.3 ± 0.76	60 ± 0	145 ± 0	70 ± 28	120 ± 87
	R	64(16)	2.1 ± 0.79	60 ± 0	145 ± 0	67 ± 27	86 ± 60
Clinical setting	L	58 (57)	2.0 ± 0.57	63 ± 9	156 ± 17	85 ± 16	85 ± 46
	R	15(15)	1.8 ± 0.64	64 ± 11	156 ± 13	75 ± 31	55 ± 23
Dizziness	L	73(39)	2.9 ± 0.77	60 ± 0	145 ± 0	−	170 ± 121
	R	30(14)	2.7 ± 0.70	60 ± 0	145 ± 0	−	136 ± 71
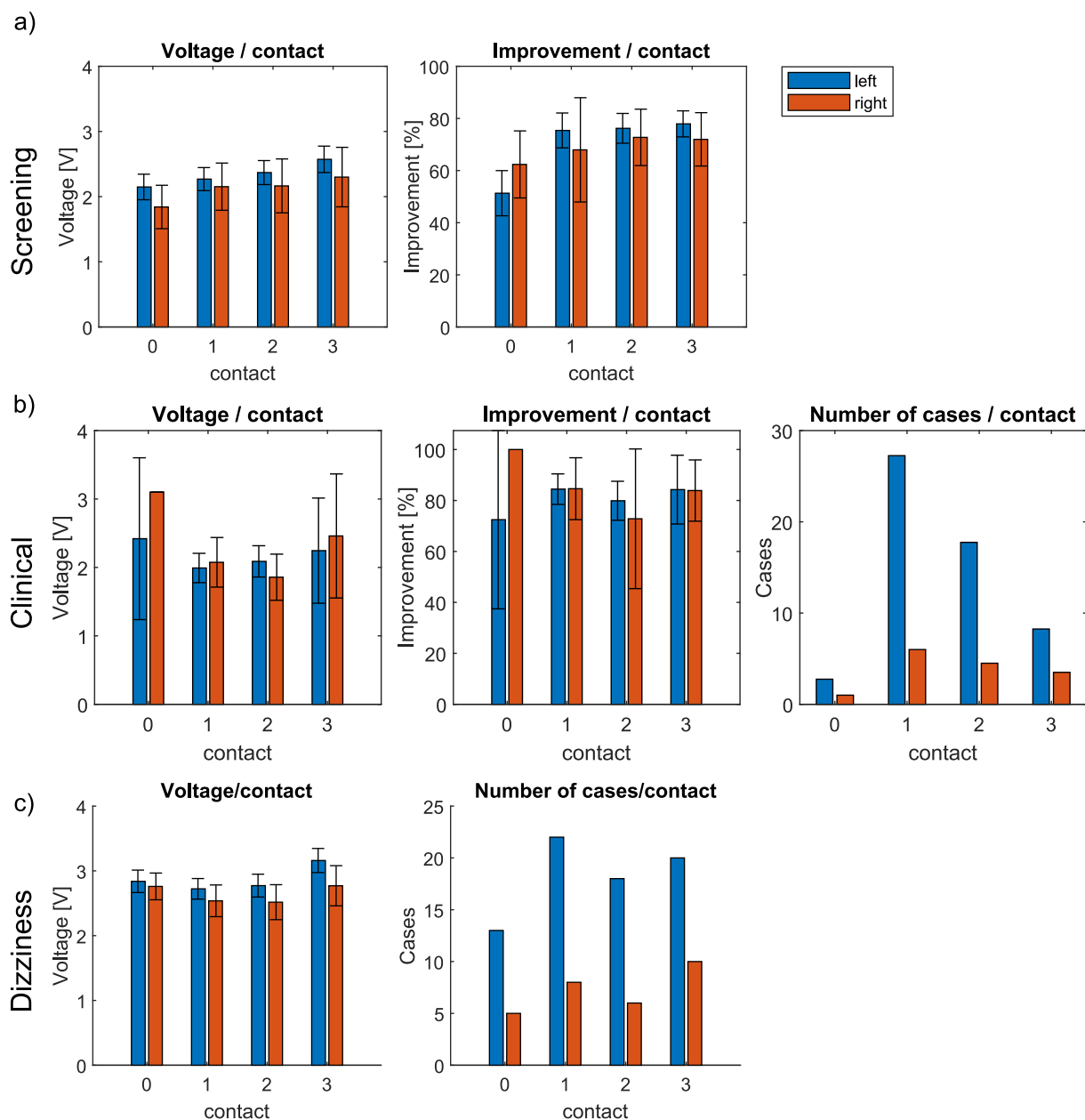
Fig. A1. The voltage and improvement for each contact are shown for a) screening, b) clinical data together with number of cases/contacts and c) dizziness. For multipolar cases in the bar charts, same improvement and voltage are assumed on all active contacts and the occurrence is divided on the active contacts. Error bars show the 95% confidence interval of the mean. For contact 0, clinical setting, the error bar was disregarded due to only 2 cases.
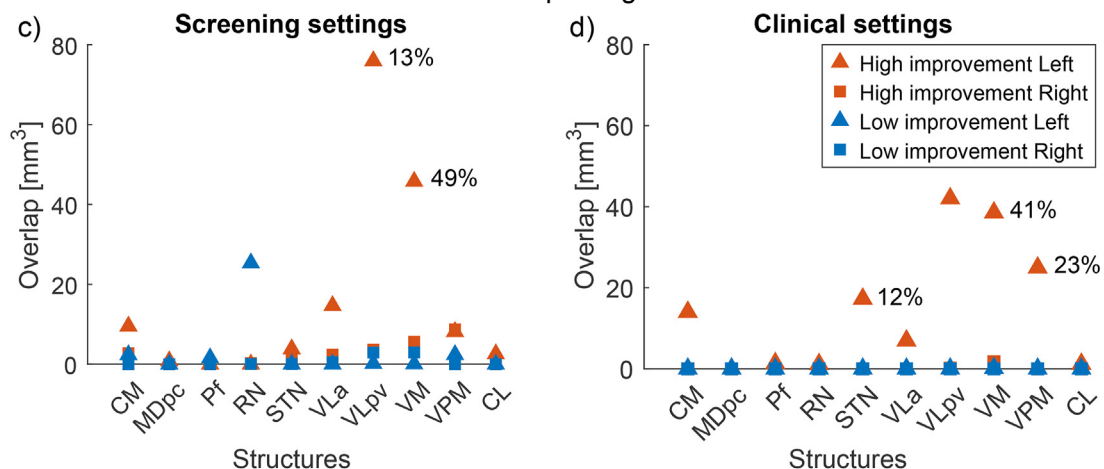
Fig. A2. Overall result from processing of clinical and screening improvement data. Visualizations are presented of a) the screening data and b) clinical data. The volumetric overlap between p-map clusters and different structures from Morels Atlas [31] for c) screening data and d) clinical data. Squares note right side and triangle left side. Blue and orange note the clusters with significantly lower and higher improvement then 70%. For overlaps >10% of the corresponding structure volume, the overlap in % are also noted.

References

[1] Chiken S, Nambu A. Mechanism of deep brain stimulation. Neuroscientist 2016;22(3):313–22. https://doi.org/10.1177/1073858415581986.

[2] Lozano AM, Lipsman N, Bergman H, Brown P, Chabardes S, Chang JW, et al. Deep brain stimulation: current challenges and future directions. Nat Rev Neurol 2019;15(3):148–60. https://doi.org/10.1038/s41582-018-0128-2.

[3] Åström M, Zrinzo LU, Tisch S, Tripoliti E, Hariz MI, Wårdell K. Method for patient-specific finite element modeling and simulation of deep brain stimulation. Med Biol Eng Comput 2009;47(1):21–8. https://doi.org/10.1007/s11517-008-0411-2.

[4] McIntyre CC, Mori S, Sherman DL, Thakor NV, Vitek JL. Electric field and stimulating influence generated by deep brain stimulation of the subthalamic nucleus. Clin Neurophysiol 2004;115(3):589–95. https://doi.org/10.1016/j.clinph.2003.10.033.

[5] Åström M, Diczfalusy E, Martens H, Wårdell K. Relationship between neural activation and electric field distribution during deep brain stimulation. IEEE (Inst Electr Electron Eng) Trans Biomed Eng 2015;62(2):664–72. https://doi.org/10.1109/TBME.2014.2363494.

[6] Elias GJB, Boutet A, Joel SE, Germann J, Gwun D, Neudorfer C, et al. Probabilistic mapping of deep brain stimulation: insights from 15 Years of therapy. Ann Neurol 2021;89(3):426–43. https://doi.org/10.1002/ana.25975.

[7] Eisenstein SA, Koller JM, Black KD, Campbell MC, Lugar HM, Ushe M, et al. Functional anatomy of subthalamic nucleus stimulation in Parkinson disease. Ann Neurol 2014;76(2):279–95. https://doi.org/10.1002/ana.24204.

[8] Dembek TA, Roediger J, Horn A, Reker P, Oehrn C, Dafsari HS, et al. Probabilistic sweet spots predict motor outcome for deep brain stimulation in Parkinson disease. Ann Neurol 2019;86(4):527–38. https://doi.org/10.1002/ana.25567.

[9] Butson CR, Cooper SE, Henderson JM, Wolgamuth B, McIntyre CC. Probabilistic analysis of activation volumes generated during deep brain stimulation. Neuroimage 2011;54(3):2096–104. https://doi.org/10.1016/j.neuroimage.2010.10.059.

[10] Stenmark Persson R, Nordin T, Hariz G-M, Wårdell K, Forsgren L, Hariz M, et al. Deep brain stimulation of caudal zona incerta for Parkinson's disease: one-year follow-up and electric field simulations. Neuromodulation: Technol. Neural. Interface 2021. https://doi.org/10.1111/ner.13500.

[11] Akram H, Sotiropoulos SN, Jbabdi S, Georgiev D, Mahlknecht P, Hyam J, et al. Subthalamic deep brain stimulation sweet spots and hyperdirect cortical connectivity in Parkinson's disease. Neuroimage 2017;158:332–45. https://doi.org/10.1016/j.neuroimage.2017.07.012.

[12] Dembek TA, Barbe MT, Åström M, Hoevels M, Visser-Vandewalle V, Fink GR, et al. Probabilistic mapping of deep brain stimulation effects in essential tremor. Neuroimage: Clinic 2017;13:164–73. https://doi.org/10.1016/j.nicl.2016.11.019.

[13] Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage 2011;54(1):313–27. https://doi.org/10.1016/j.neuroimage.2010.07.033.

[14] Vogel D, Shah A, Coste J, Lemaire J-J, Wårdell K, Hemm S. Anatomical brain structures normalization for deep brain stimulation in movement disorders. Neuroimage: Clinic 2020;27:102271. https://doi.org/10.1016/j.nicl.2020.102271.

[15] Åström M, Samuelsson J, Roothans J, Fytagoridis A, Ryzhkov M, Nijlunsing R, et al. Prediction of electrode contacts for clinically effective deep brain stimulation in essential tremor. Stereotact Funct Neurosurg 2018;96(5):281–8. https://doi.org/10.1159/000492230.

[16] Wårdell K, Nordin T, Vogel D, Zsigmond P, Westin C-F, Hariz M, et al. Deep brain stimulation: emerging tools for simulation, data analysis, and visualization. Front Neurosci 2022;16:834026. https://doi.org/10.3389/fnins.2022.834026.

[17] Dembek TA, Baldermann JC, Petry-Schmelzer J-N, Jergas H, Treuer H, Visser-Vandewalle V, et al. Sweetspot mapping in deep brain stimulation: strengths and limitations of current approaches. Neuromodulation: Technol. Neural. Interface 2021. https://doi.org/10.1111/ner.13356.

[18] Fytagoridis A, Åström M, Samuelsson J, Blomstedt P. Deep brain stimulation of the caudal zona incerta: tremor control in relation to the location of stimulation fields. Stereotact Funct Neurosurg 2016;94(6):363–70. https://doi.org/10.1136/jnnp-2017-317219.

[19] Fytagoridis A, Sandvik U, Åström M, Bergenheim T, Blomstedt P. Long term follow-up of deep brain stimulation of the caudal zona incerta for essential tremor. J Neurol Neurosurg Psychiatr 2012;83(3):258. https://doi.org/10.1136/jnnp-2011-300765.

[20] Blomstedt P, Sandvik U, Tisch S. Deep brain stimulation in the posterior subthalamic area in the treatment of essential tremor. Mov Disord 2010;25(10):1350–6. https://doi.org/10.1002/mds.22758.

[21] Fahn S, Tolosa E, Marín C. Clinical rating scale for tremor. In: Jankovic J, Tolosa E, editors. Parkinson's disease and movement disorders. Baltimore-Münich: Urban and Schwarzenberg; 1988. p. 225–34.

[22] Wårdell K, Zrinzo L, Hariz M, Andersson M. Patient-specific brain modelling for deep brain stimulation simulations. International IEEE/EMBS Conference on Neural Engineering (NER); 2013. p. 148–51. 2013 6th.

[23] Alonso F, Hemm-Ode S, Wårdell K. Influence on deep brain stimulation from lead design, operating mode and tissue impedance changes - a simulation study. Brain Disord Ther 2015;4(3):169. https://doi.org/10.4172/2168-975X.1000169.

[24] Nielsen MS, Bjarkam CR, Sørensen JC, Bojsen-Møller M, Sunde NA, Østergaard K. Chronic subthalamic high-frequency deep brain stimulation in Parkinson's disease – a histopathological study. Eur J Neurol 2007;14(2):132–8. https://doi.org/10.1111/j.1468-1331.2006.01569.x.

[25] Johansson JD, Zsigmond P. Comparison between patient-specific deep brain stimulations and commercial system SureTune3. Biomed Phys Eng Express 2021;7(5). https://doi.org/10.1088/2057-1976/ac0dcd.

[26] Kirsch AD, Hassin-Baer S, Matthies C, Volkmann J, Steigerwald F. Anodic versus cathodic neurostimulation of the subthalamic nucleus: a randomized-controlled study of acute clinical effects. Park Relat Disord 2018;55:61–7. https://doi.org/10.1016/j.parkreldis.2018.05.015.

[27] Kuncel AM, Cooper SE, Grill WM. A method to estimate the spatial extent of activation in thalamic deep brain stimulation. Clin Neurophysiol : Off. J. Int. Fed. Clin. Neurophysiol. 2008;119(9):2148–58. https://doi.org/10.1016/j.clinph.2008.02.025.

[28] Vogel D, Wårdell K, Coste J, Lemaire J-J, Hemm S. Atlas optimization for deep brain stimulation. In: Jarm T, Cvetkoska A, Mahnič-Kalamiza S, Miklavcic D, editors. 8th European medical and biological engineering conference. Cham: Springer International Publishing; 2021. p. 130–42.

[29] Ewert S, Horn A, Finkel F, Li N, Kühn AA, Herrington TM. Optimization and comparative evaluation of nonlinear deformation algorithms for atlas-based segmentation of DBS target nuclei. Neuroimage 2019;184:586–98. https://doi.org/10.1016/j.neuroimage.2018.09.061.

[30] Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, et al. The optimal template effect in Hippocampus studies of diseased populations. Neuroimage 2010;49(3):2457. https://doi.org/10.1016/j.neuroimage.2009.09.062.

[31] Morel A. Stereotactic atlas of the human thalamus and basal ganglia. CRC Press; 2007.

[32] Jakab A, Blanc R, Berényi EL, Székely G. Generation of individualized thalamus target maps by using statistical shape models and thalamocortical tractography. Am J Neuroradiol 2012;33(11):2110. https://doi.org/10.3174/ajnr.A3140.

[33] Nowacki A, Schober M, Nader L, Saryyeva A, Nguyen T-AK, Green AL, et al. Deep brain stimulation for chronic cluster headache: meta-analysis of individual patient data. Ann Neurol 2020;88(5):956–69. https://doi.org/10.1002/ana.25887.

[34] Reich MM, Horn A, Lange F, Roothans J, Paschen S, Runge J, et al. Probabilistic mapping of the antidystonic effect of pallidal neurostimulation: a multicentre imaging study. Brain 2019;142(5):1386–98. https://doi.org/10.1093/brain/awz046.

[35] Nguyen TAK, Nowacki A, Debove I, Petermann K, Tinkhauser G, Wiest R, et al. Directional stimulation of subthalamic nucleus sweet spot predicts clinical efficacy: proof of concept. Brain Stimul 2019;12(5):1127–34. https://doi.org/10.1016/j.brs.2019.05.001.

[36] Liewald D, Miller R, Logothetis N, Wagner H-J, Schüz A. Distribution of axon diameters in cortical white matter: an electron-microscopic study on three human brains and a macaque. Biol Cybern 2014;108(5):541–57. https://doi.org/10.1007/s00422-014-0626-2.

[37] von Keyserlingk Graf D, Schramm U. Diameter of axons and thickness of myelin sheaths of the pyramidal tract fibres in the adult human medullary pyramid. Anat Anzeiger 1984;157(2):97–111.

[38] Hemm S, Mennessier G, Vayssiere N, Cif L, El Fertit H, Coubes P. Deep brain stimulation in movement disorders: stereotactic coregistration of two-dimensional electrical field modeling and magnetic resonance imaging. J Neurosurg 2005;103(6):949–55. https://doi.org/10.3171/jns.2005.103.6.0949.

[39] Horn A, Li N, Dembek TA, Kappel A, Boulay C, Ewert S, et al. Lead-DBS v2: towards a comprehensive pipeline for deep brain stimulation imaging. Neuroimage 2019;184:293–316. https://doi.org/10.1016/j.neuroimage.2018.08.068.

[40] Kübler D, Kroneberg D, Al-Fatly B, Schneider G-H, Ewert S, van Riesen C, et al. Determining an efficient deep brain stimulation target in essential tremor - cohort study and review of the literature. Park Relat Disord 2021;89:54–62. https://doi.org/10.1016/j.parkreldis.2021.06.019.

[41] Mai JK, Majtanik M, Paxinos G. Atlas of the human brain. Academic Press; 2015.

[42] Shah A, Coste J, Lemaire J-J, Taub E, Schüpbach WMM, Pollo C, et al. Intra-operative acceleration measurements to quantify improvement in tremor during deep brain stimulation surgery. Med Biol Eng Comput 2017;55(5):845–58. https://doi.org/10.1007/s11517-016-1559-9.

[43] Nowacki A, Barlatey S, Al-Fatly B, Dembek T, Bot M, Green AL, et al. Probabilistic mapping reveals optimal stimulation site in essential tremor. Ann Neurol 2022;91(5):602–12. https://doi.org/10.1002/ana.26324.