Appendix: Approaches for determining Sweetspots in Deep Brain Stimulation

This appendix first outlines the general concepts of mapping stimulation volumes (VTAs) in deep brain stimulation (DBS). We then provide a detailed overview about the different methods for determining DBS sweetspots which have been investigated for this study. Additionally, we offer theoretical considerations for each method, which we deemed to be important for users.

Prerequisites for DBS mapping

To realize a DBS mapping project over a cohort of patients, a certain set of prerequisites have to be met. First, some method of relating the stimulation to a neuroanatomical frame of reference needs to be established. First the locations of the electrodes need to be determined from postoperative imaging. Then most studies use the concept of a binary stimulation volume, the VTA to assess which parts of the brain have either been stimulated or not with a certain set of stimulation parameters. The VTA itself is of course a simplification and many different VTA-models as well as different VTA thresholds, or underlying brain conductivities have been suggested, all influencing the respective size and shape of a VTA. Instead of using VTAs, alternative concepts may include using the electrode coordinates and a weighting function based on the distance to the active electrode or for example weighting stimulation influence using a non-thresholded electric field.

A second prerequisite is that results from individual patients are transferred to a common neuroanatomical reference space, either by e.g. nonlinear normalization of preoperative MRI images or using landmarks.

Once these prerequisites have been met, a variety of strategies can be employed to determine DBS sweetspots. In the following, we will explain the basics of DBS mapping. Then we will describe in detail the five methods investigated for this study and suggest some theoretical considerations regarding the properties of each method.

DBS mapping basics

Two sets of basic images, namely the N-image and the Mean-image (also called Mean-Effect-Image), serve as the basis of all mapping strategies investigated in this study and thus need to be explained in detail (see Figure S1). The N-image is generated by summation of all binary VTAs within the common reference space. Each resulting voxel of the N-image thus provides the information of how many times it was covered by VTAs of the underlying dataset. Some mapping strategies rely on the N-image itself to define their sweetspot, while others use it for discarding voxels which have not been stimulated enough times and as the basis for further analysis. Due to the nature of its generation, the center of the N-image typically shows the largest values, since it is there that the most VTAs overlap. The Mean-image on the other hand, incorporates not only the information of where stimulation occurred but also includes information about the clinical outcomes of the respective stimulation settings. To achieve this, binary VTAs are `clinically-informed` by assigning them their respective clinical outcomes, in general via multiplication of the VTAs voxels with the improvement scores. By again summatining these clinically-informed VTAs and then dividing the resulting image by the N-image each voxel in the resulting Mean-image contains the average improvement of all VTAs covering this voxel. Due to the fact, that most VTAs overlap in the center of the Mean-image, resulting in a more general average of outcomes in these voxels, while fewer and
larger VTAs (often associated with more extreme outcomes) overlap at its fringes, the Mean-image has an inherent tendency to overemphasize its outer parts and underemphasize its center.

**Figure S1 Legend:** Example of how the N-image and the Mean-image are generated. The sensorimotor STN is shown in blue (3d) and as outline (2d). The N-image is created by calculating the number of times each voxel is covered by a VTA. For the Mean-image, clinically informed VTAs are generated by assigning each voxel of each VTA the corresponding clinical outcome. Then, for each voxel in the reference space an average outcome is calculated by averaging all clinical outcomes of all VTAs covering this voxel.
In 2011 Butson et al. published “Probabilistic Analysis of Activation Volumes Generated During Deep Brain Stimulation”. In this study the authors used monopolar review data from six Parkinson’s disease (PD) patients which had been implanted in the subthalamic nucleus (STN). Overall, their dataset consisted of 163 stimulation settings and outcomes included motor symptoms like bradykinesia and rigidity.

Mapping method

For their mapping method Butson and colleagues generated a Mean-image from VTAs for all 163 stimulation settings as described above. They then applied a threshold to select only voxels which showed at least an average improvement of 50 % in the Mean-image. Additionally, they repeated the same analysis with a threshold of 75 % which of course lead to a smaller sweetspot. For our comparison of methods, the 50 % threshold was used, since it resulted in sweetspot sizes closer to the ground-truth.

Theoretical considerations

As explained above, the Mean-image tends to show larger differences at its outer fringes due to a variety of factors:
1. Most VTAs overlap in the center of the Mean-image, so that values in the center of the Mean-image often resemble a general average of outcomes.
2. Fewer VTAs overlap at its outer fringes, making those voxels in the Mean-image more susceptible to outliers
3. Especially in monopolar review data, outer voxels are mainly stimulated by VTAs generated by larger stimulation amplitude. Larger amplitudes however also correlate with higher improvement during monopolar reviews, resulting in higher average improvement scores at the outer fringes if data is not corrected for stimulation amplitude.

Figure S2 Legend:
Schematic illustration of the Butson et al. method. The sweetspot is highlighted as red outline.
Cheung et al. 2014 – Thresholding the N-image for responders in clinical DBS settings

In 2014 Cheung et al. published “Defining a Therapeutic Target for Pallidal Deep Brain Stimulation for Dystonia”. In their study, they investigated the effects of clinically used stimulation settings in 21 patients with dystonia who had been implanted into the pars internus of the Globus pallidus (GPi).

Mapping method

For each of the \( n = 42 \) leads the corresponding VTA was calculated. Then the 25% VTAs with the worst clinical outcomes were discarded. From the remaining 75% of VTAs (\( n = 32 \)) an N-image was created as described above. To define their sweetspot, the authors then selected only those voxels from the N-image, which were covered by at least 75% of the \( n = 32 \) VTAs.

As discussed above, all N-images show their largest value at the center, where most VTAs overlap. Sweetspots based on thresholded N-images, thus will always lie at the center of the investigated VTAs and can never lie outside of the coordinates where most leads are implanted. Such method will thus not identify ‘off-target’ sweetspots. Limiting the analysis to responders can of course move the N-image towards the ‘true’ sweetspots. However, the neuroanatomical information of non-responders, might be equally important in determining a sweetspot, but is lost in this approach.

Figure S3 Legend:
Schematic illustration of the Cheung et al. method. The sweetspot is highlighted as red outline.
Eisenstein et al. 2014 – Introducing voxel-wise statistics and mapping non-motor outcomes

In their landmark publication Eisenstein et al. were the first to introduce voxel-wise statistics into the field of DBS mapping. Furthermore, in their publication about motor and non-motor effects of STN-DBS they also were the first in the field to address the problem of multiple comparisons, by also incorporating an adequate post-hoc analysis, in their case a nonparametric permutation statistic. Their dataset consisted of clinical stimulation settings in 51 PD patients and a variety of motor, cognitive, and emotional parameters.

Mapping method

Interestingly, the Eisenstein et al. method did not use stimulation parameters or a VTA-model. Instead it assigned clinical values to voxels by weighting the clinical outcomes with the distance from the stimulating electrode. They based their weighting on an estimate that stimulation voltage would drop by 50% over a distance of 1.5 mm in gray matter. By applying a 3D Gaussian function based on this decrease, they could then weight clinical outcomes in each voxel depending on the distance to the center of the active electrode – eliminating scores with a weighting of less than 5%.

This data was then used to create Mean-Images as previously described. Then a two-sided t-test was used to compare the weighted outcome values in each voxels against a null-hypothesis of zero (no change in outcome). The corresponding p-value of each voxel-wise t-test was stored in the \( p \)-Image. Voxels with a significant difference from zero (\( p < 0.05 \)) were identified.

**Figure S4 Legend:** Schematic illustration of the Eisenstein et al. method. The sweetspot is highlighted as red outline.
To address the issue of multiple comparisons, the authors used a nonparametric permutation statistic to analyze the overall statistical significance of their p-Image. As marker of the overall statistical significance, they calculated the sum of negative decadic logarithms of each significant p-value ($p < 0.05$) in the p-Image. This summary statistic was then compared against summary statistics from 200 permuted datasets, which were generated the same way as the original dataset but after randomly pairing the clinical outcomes of one patient with the stimulation locations from another patient. The statistical results in the original dataset were only deemed valid, if their summary statistic was larger than in 95% of the 200 permuted datasets.

Theoretical considerations

The approach by Eisenstein et al. greatly expanded the field of DBS mapping by moving from a purely descriptive view and simple thresholding towards a statistical validation of mapping results. As the authors describe themselves, three limitations need to be considered:

1. Since no VTA-model was used, differences in stimulation parameters do not impact the results of their mapping approach. On the one hand this is a limitation, since different parameters of course have different effects in a given patient. On the other hand, many of the underlying properties of the VTA (e.g. conductivity, activation) rely on insufficiently defined values/thresholds. Avoiding VTAs as a concept thus might also be considered a strength.

2. Performing statistical tests against zero as a null-hypothesis might lead to a lot of positives in case of one-sided clinical outcomes like change in motor outcome. Since motor symptoms in most cases tend to improve and almost never worsen due to stimulation, this approach might lead to large areas of significant voxels with a low spatial specificity.

3. The summary statistic employed during the nonparametric post-hoc analysis is only based on the overall significance of the p-Image as a whole. Thus it neglects the topography of the p-Image including the spatial distribution of significant voxels into e.g. clusters.
Reich et al. 2019 – Connected VTAs versus unconnected VTAs

Reich and colleagues used sweetspot mapping to define an optimal stimulation target in a dataset of chronic DBS settings in 87 dystonia patients. Additionally to calculating a binary sweetspot, they also used VTA-overlaps with the underlying statistical maps (T-images) to predict outcome in a leave-one-out fashion and in an independent dataset.

Mapping method

Analogous to Butson et al., Mean-images were created by assigning each binary VTA the corresponding clinical outcome. The authors then compared all the clinical outcomes, associated with the stimulation of one voxel, to all the remaining outcomes of stimulation settings not reaching that particular voxel, using a two-sided t-test.

After this voxel-wise, statistical analysis, the sweetspot was defined by selecting only voxels with p-values $p < 0.05$, and by furthermore discarding all voxels which did not belong to spatial clusters of at least 500 voxels.

Theoretical considerations

Statistically comparing the outcomes of VTAs stimulating one voxel to those VTAs not stimulating that particular voxel addresses the problem observed in Eisenstein et al. by defining a more conservative null-hypothesis. The benchmark for statistical significance is thus moved from “voxel does lead to improvement” to “voxel leads to exceptional improvement”. This, at least in theory, should focus the resulting sweetspot on those areas being associated with the best possible outcome observed in a dataset.

The arbitrarily defined cluster threshold as a post-hoc protection against false positives, is of course strongly influenced by underlying factors like the voxel-size used for analysis and thus might not be universally adequate.
Dembek et al. 2019 – Mapping monopolar reviews and correcting for amplitude

Dembek and colleagues tried to determine sweetspots for parkinsonian motor symptoms using their DBS mapping method in prospective monopolar review data. They could demonstrate, that their sweetspots were able to predict outcome in one independent dataset of other monopolar review data, but less so in another independent dataset with chronic stimulation settings.

Mapping method

Mean-images were created by assigning each binary VTA the corresponding clinical outcome as described before. The authors then compared each outcome, associated with the stimulation of one voxel, to all other outcomes which stemmed from stimulation at the same amplitude using voxel-wise Wilcoxon signed-rank tests. As in Reich et al., sweetspots were defined by selecting only voxels with p-values $p < 0.05$.

As suggested by Eisenstein et al., a non-parametric permutation analysis was performed to address the issue of multiple comparisons. Contrary to Eisenstein et al., summary statistics were not calculated on the whole image, but for each cluster of significant voxels separately. Thus individual clusters were accepted or discarded when their summary statistic was larger/smaller than the summary statistics of the $95\%$ highest ranking clusters found in 1000 datasets generated from permuted outcomes.

Theoretical considerations

Comparing each clinical outcome against the average outcome of stimulations with the same amplitude signifies an amplitude-correction based on the clinical outcomes of the dataset itself. In monopolar review data, this is a reasonable approach, since a strong relationship between outcomes and amplitudes can be observed. When stimulating on an electrode, even a suboptimally placed one, with increasing amplitudes a positive relation between outcome and amplitude is very likely to occur and thus needs to be addressed.

On the other hand, it remains to be seen whether such a strategy is also valid when using the method on with chronic DBS settings. Chronic DBS settings often are far more complex than the clearly defined settings during monopolar reviews. Consecutively, comparing the outcomes of one setting to “equal” settings is much less straightforward. Additionally, patients with a well-placed electrode might experience optimal outcome at low stimulation strengths while patients with off-target electrodes might need higher stimulation amplitudes. The relationship between stimulation amplitude and outcome observed during monopolar reviews might thus be inexistent or even be an inverse relationship in chronic DBS data.
Mapping methods:


Further references:


