Pre-registered report: Space Sequence Synesthesia Diagnostic using form mapping

Rémy Lachelin, Chhavi Sachdeva, and Nicolas Rothen

Psychology, UniDistance Suisse

Author Note

Rémy Lachelin https://orcid.org/0000-0002-8485-7153

Chhavi Sachdeva https://orcid.org/0000-0002-0074-4371

Nicolas Rothen https://orcid.org/0000-0002-8874-8341

Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: Rémy Lachelin: design, analyses; Chhavi Sachdeva: design, data collection, manuscript; Nicolas Rothen: design, founding, manuscript

Correspondence concerning this article should be addressed to Rémy Lachelin, Psychology, UniDistance Suisse, Schinerstrasse 18, Brig-Glis, Valais 3900, Email: remy.lachelin@fernuni.ch

Abstract

Sequence-space synesthesia (SSS) is the phenomenon of representing ordered visual symbols in particular spatial positions. Existent tools to detect SSS are based on self-reports (i.e. questionnaire) and consistency tests. Consistency tests are critical for the understanding of synesthesia to characterize or classify participants' synesthesia in experimental settings. We attempt to further optimize SSS diagnostic criteria with a paradigm shift. In this pre-registered report, we compare available diagnostics with new diagnostic criteria on 685 participants. Conceptually, the novel criteria aim at taking advantage of ordinality and extract geometric features based at the forms level. We harness a geography package to compute new potential criteria. Receiver Operator Characteristics analyses are used to compare all features. The results suggest topological validity to be the best criteria. In a second phase, we will test the predictive power of the new diagnostic features on an additional dataset that has yet to be collected.

Keywords: Space sequence synesthesia, consistency test

Pre-registered report: Space Sequence Synesthesia Diagnostic using form mapping

Introduction

Sequence-Space Synesthesia (SSS) or visuo-spatial forms is the phenomenon where people visualize ordered sequences in particular spatial positions. For example, numbers, weekdays or months (synesthetic *inducers*) are represented as arranged into specific spatial positions is space (synesthetic *concurrent*).

Heterogeneity and homogeneity of Sequence-Space Synesthesia

The spatial (visuo-spatial) forms of SSS are idiosyncratic, which results in considerable heterogeneity in how this phenomenon is manifested across individuals. One source of heterogeneity is given by dimensionality: some SSS experiences involve three-dimensional (3D) and two-dimensional (2D) spatial arrangement (Eagleman, 2009; Price & Pearson, 2013). Another source is the reference frame, for example , the spatial forms take place in an external space around the body (*i.e.* projector) or in an internal space (*i.e.* associator) (Dixon et al., 2004; Smilek et al., 2007). Further variability can be explained by temporal-spatial properties, such as manipulation of their spatial forms such as "zooming" in and out, rotating or shifting perspectives (Gould et al., 2014). Lastly, the shape, complexity and layout of the spatial forms are also heterogeneous such as forming for example ovals, lines or zig-zags or loops. With some recurring shapes being more frequent, such as ovals for months (Eagleman, 2009).

Despite the after mentioned heterogeneities, SSS is also phenomenological characterized (Seron et al., 1992). *Automaticity*: the *inducer* automatically triggers the *concurrent*. *Unidirectionality*: while the *inducer* triggers the concurrent, the concurrent does not trigger the inducer. *Developmentally early*: the experience was already present during childhood. *Consciousness*: The concurrent is consciously perceived. *Consistency*: the inducer-concurrent pair remains stable in time within subject. These distinctions can be quantified with self reported questionnaire (i.e. for development and consciousness), or more objectively using behavioural tests such as consistency tests (Baron-Cohen et al., 1993).

Consistency tests

The rationale behind consistency tests are designed to measure the variability in *inducer-concurrent* across time. These test are used as an objective validation or genuinnes test of self-reported synesthetes, it is therefore mainly useful in experimental settings to compare synesthetes and control and characterize the former.

Consistency tests have made their proof for colour-grapheme synesthesia. Measures of individual consistency can be derived using colour-pickers while presenting an inducer repeatedly within a list (i.e. the letter "A"). More specifically the euclidean distance in CIE-LUV colour space (which is designed for perceptual uniformity) between repetitions of the same inducer lead to satisfactory accuracy when using a cut-off estiated from a larger sample (Rothen et al., 2013).

A similar rationale than for colour-grapheme synesthesia has been used to characterize SSS. Brang et al.(2010), evaluated consistency as the distance between repetition (i.e. January and January) compared to the adjacent stimuli (i.e. February), the stimulus response was defined as consistent if within 1.96 z-scores. This criteria was however noted to potentially be to conservatory, since it detected synesthesia in 4 of 81 self-reported synesthetes.

The same rationale as for colour-grapheme synesthesia (Rothen et al., 2013) has been applied to design a consistency test of SSS (Rothen et al., 2016a). For SSS, instead of a colour-picker, participant's task it to position each selected inducer on the computer's screen position of their concurrent experience with each inducer being usually repeated three times. Form this task, the area and perimeter between the coordinates of repeated inducers has been used as a measure of consistency (Rothen et al., 2016a). Consistent SSS responses should lead to smaller area (i.e. closer response in space). A cut-off of an average triangle area covering <.203 % of the total screen area to classify as SSS was suggested i.e. 1596 pixels on a 1024 X 768 resolution display (see also Ward et al., 2018). Standard deviation of responses and permuting the responses to compare single responses to a chance level calculated from permutations as in Root (2021) have been compare in (Ward, n.d.-a). A measure of consistency between the x and y coordinates is then compared. The total area between the responses of same inducer has been

suggested to be used (Rothen et al., 2016a).

(Ward et al., 2018)

One general caveat for consistency tasks, is that synesthetic forms are idiosyncratic. In other words, the inducer-concurrent pairs might lead to form. Other problems with this type of consistency tests is that (1) it might favor one form of SSS, such as those with linear spatial forms (Ward, n.d.-a). (2) some participant not knowing the responses might click on the same position, leading to high consistencies (see Rothen et al., 2016a). (3)

Moreover, this kind of criteria might bias the diagnosis to include synesthetes with straight lines which leads to less variability than more complex forms(**?**).

Present study

The goal of this registered report is to compare different consistency features on their ability to classify SSS from controls using Receiver Operator Characteristics (ROC). In the present *Phase I* we merge already available datasets to replicate systematic consistency test methods from the literature to additional new features. These new features are designed to take advantage of two properties of synesthetic responses First, the importance of ordinality between the inducers (i.e. Monday -> Tuesday -> Wednesday, ect). Some studies have systematically investigated ordinality, but using adjacent inducers (i.e. the distances between Monday, Tuesday and Wednesday as in (Brang et al., 2010). Other have used the angle formed by adjacent inducers (Eagleman, 2009).

Second, thee particular synthetic forms of the sequential spatial location. These forms might have geometrical properties. For example months of the year might be represented circularly (as already described by (Galton, 1880) for numbers).

To take advantage of sequential and geometrical synesthetic forms, we harnessed a geo-spatial package(Pebesma, 2018) to extract geometrical features from participant x and y coordinate responses. This packages allows for example to build string or polygons for each repetition and compare different geometrical features. Those individual geometrical features are then compared using Receiver Operator Characteristics (ROC) between individuals grouped as synesthets and control. In the present *phase I*,we compare ROC on three merged derivation

datasets using the same task on SSS Ward (n.d.-a). In future *phase II*, we compare whether the features selected to diagnose SSS in *phase I,* on a validation dataset that is not yet acquired (registered report on the open science foundation: https://osf.io/9efjb/**).**

The rationale here is that synesthetic responses should have geometrical feature that differ from controls. For example, several SSS representations for months are circular.

General Methods

Phase I: present analyses. We merged three available datasets and compared available diagnostic criteria across datasets using Receiver Operator Characteristics (ROC) for different approaches. First, we attempt at reproducing the diagnostic criteria on stimulus level consistency such as area and perimeter. Second, we explore a new approach consisting of comparing geometrical features across repetitions. Third, we apply second order approaches such as permutation tests. On one hand we reproduce Root et al. (2021) permutation test that was developed for colour-grapheme synesthesia applied to SSS, as in ward (n.d.-a)). On the other hand we apply permutations across repetitions to obtain a permuted measure of consistency. In other words, we shuffle the presentation order and compute the geometrical feature in non chronological order. Finally we use General Linear Model's (GLM) to attempt to fit the best diagnostic curve combining different criteria.

Phase II: future analyses. On a future dataset to be collected using the same task, we will compare the predictive power of the selected features using ROC.

Materials

A the exception of (Rothen et al., 2016a) (see https://osf.io/6hq94/files/osfstorage), the data from (Van Petersen et al., 2020a; Ward, n.d.-a) were collected online. The 29 inducers were: the 12 months of a year, 7 days of the week and 10 numbers (i.e. hindo-arabic numerals from 0 to 9). (Van Petersen et al., 2020a) Also presented 50 and 100 numerals, which we excluded here. (Ward, n.d.-a) data is collected using the Syntoolkit.

Procedure

The details for the task's administration of each dataset are described in each respective article: (Ward, n.d.-a) conducted the task online and (Rothen et al., 2016a; Van Petersen et al., 2020a) in laboratory. Each stimuli is presented randomly and sequentially centrally on the screen. Participant are instructed to click on the screen position where they visualize them.

Stimulus included here are 7 weekdays (Monday to Sunday), 12 months (January to December) and 9 numbers (0 to 9). For Ward's data the stimulus were presented in randomized order with the constraint that no stimulus was repeated until all unique stimuli (N = 29) had been presented once.

Importantly, while the three datasets included in *phase I* include three repetitions per stimuli, the *phase II* will use four repetition per stimuli.

Phase I. Methods

Phase I. Participants

We merged four datasets: Rothen et al. (2016a),(Ward, n.d.-a) (from: https://osf.io/p5xsd/files/osfstorage) ,(Van Petersen et al., 2020b) and additional data gently provided by private communications with Pr. Ward, see Table 2. To match the other datasets, stimuli form (Van Petersen et al., 2020b) are translated from Dutch to English and for the stimuli, only numbers from 0 to 9 are kept (excluding 50 and 100).

We kept 685 from the initial 689 participants. First, we excluded 1307 empty trials including trials flagged for having the same x or y coordinates across conditions and repetitions, causing the depletion of 2 participants(as in Rothen et al., 2016b; Ward et al., 2018). 2 participants were excluded for having less than 4 coordinate points since this would impeach computing polygons, see Section . As a consequence, 0 participants did not have coordinates in all conditions, for example no coordinates for numbers. Then, x and y coordinates were then separately normalized (z-score) per participant.

From the final sample of N = 685, 396 were synesthetes and 289 controls. Table 2 breaks down the number of synesthetes and control contributed by each dataset.

Since not all the data is directly associated to demographics, we resume in Table 1 gthe original sample's reported descriptives.

Regarding the synesthes, we can only describe their profile from the data by Pr. Ward (i.e. from 573 cases. These profiles are described only for the stimulus class that are used in the consistency test (i.e. number, weekdays and month), see Figure 1.

Phase I. Procedure

The median display resolution was 1440 X 768, with a maximum of 2560 X 2025 and a minimum of 308 X 149 .

Phase I. Analysis

First, we reproduce consistency methods found in the literature using the same task ((Root et al., 2021; Rothen et al., 2016a; Van Petersen et al., 2020a; Ward, n.d.-a)) and compare the results. These methods are on the stimulus level, hence they asses the consistency for each stimulus *within* the repetitions.

Second, we extract features on the form level. We harness a geography package to compute geometry based features. Informed on the ordinality of the stimulus (i.e. monday, tuesday, ect), we construct segments and polygon by conditions and repetitions. These methods are on the form or category level. The rationale here is to see whether when considering the stimuli as ordered coordinates, i.e. as segments or polygon, they are consistent *between* repetitions.

Since these methods are also relying on repetition order (i.e. the segment for numbers are constructed with repetition 1 - vs. 2 vs. 3, that is their chronological order of experimental presentation), we also compute the best AUC features by permuting repetitions. We predict that permuted averaged features should lead to better classifications. Because synesthete's within stimulus consistency should lead to similar forms interdependently from the chronological order of stimulus presentation.

Finally, we also intent a corelational approach.

Stimulus level: area and perimeter between repetitions

Conceptually, the more consistent responses to the same stimuli should have smaller coordinate distance. This distance can be computed as the area or the perimeter formed by the x,y coordinate between the repetitions (i.e. (x1, y1), (x2, y2), (x3, y3)). With three repetitions, the area is calculated as in using the formula Equation 1 and the perimeter Equation 2.

$$Area = (x1y2 + x2y3 + x3y1x1y3x2y1x3y2)/2 \tag{1}$$

$$Perimeter = \sqrt{(x2 - x1)^2 + (y2y1)^2} + \sqrt{(x3 - x2)^2 + (y3y2)^2} + \sqrt{(x1 - x3)^2 + (y1y3)^2} \tag{2}$$

We compute the area in term of % of the screen size to be able to compare with the consistency results in (Rothen et al., 2016a) and (Ward et al., 2018). In addition, since the screen sizes differ between experimental settings and individual response's spread vary, we computed the area on individually z-score transformed x,y coordinates. The area of each stimuli is then averaged for each participants.

Next we replicate (Root et al., 2021) permuted consistency method. For each individual, the x and y coordinates are randomly permuted and the areas are calculated. After 1000 permutations per individual, a z-score is calculated with the observed means compared to the permuted distribution, see Equation 3.

$$Zscore = [(Observed)(MeanPermuted)]/(SDPermuted) \tag{3}$$

The permuted distribution would be the theoretical individual chance level distribution and hence the z-score reflects where the observed area lies from the chance level.

Category level: additional features

Taking the stimuli as an ordered sequence we can consider them as a geometrical segments (i.e. open geometrical form) and polygons (i.e. closed geometrical form), similarly as originally described in Galton (1880). From there we can extract several geometrical properties.

First, we extract the number of *self-intersections* of each segments. Conceptually, SSS should have less chance to produce that self-intersect than control. The number of self-intersections are added separately for each repetitions and conditions and averaged per participants.

The next geometrical features are extracted using the simple feature `sf` package (Pebesma, 2018) to generate ordered segments and polygons based on the individually z-score transformed x and y coordinates. `sf` has originally been developed for geography.

We calculate the polygon's area and perimeter of the polygons and average per participants.

Then we use the second order functions provided by the **`sf`** package to check for topological features from the polygons. These tests return a boolean. First we test each polygon for simplicity, a simple polygon being described as not having self-intersections or self-tangencies.

Then we test for topological validity, this function tests if a polygon is *is well-formed and valid in 2D according to the* Open Geopsatial Consotrtium *rules (see* https://postgis.net/docs/ST_IsValid.html and https://postgis.net/docs/using_postgis_dbmanagement.html#OGC_Validity).

Finally, we also attempted a correlational approach. Here we correlate the coordinates.

Category level: additional improvements

Until now the form based features are computed by chronologically ordered repetitions. For example, Monday is repeated three times per ID. The coordinates for Monday presented the first time will always be used to form the segment/polygon with the Tuesday presented the first time. However, for consistency, this should be independent from chronological order. To circumvent this, we can permute the repetitions per conditions. I predict the permuted averages of the same features should give rise to better AUC.

Finally, we used General Linear Model (GLM) on the two features with the best AUC and add the prediction as an additional feature. A GLM could provide a formula where multiple features could be combined in order to optimize AUC.

Phase I. Results

The average values for each features are reported in Table 3. For the area between repetitions, we find different areas for synesthetes than in previous results: 0.26% , compared to 0.14% in (2016a) and 0.15 % in (Ward et al., 2018)). Note that this difference seems to be mainly driven by the dataset from (Ward, n.d.-a) with respectively 0.26% area. Descriptively, this might be explained by a more variable sample [SD of 0.50 in (n.d.-a) compared to 0.09 % in Rothen et al. (2016a)] see Table B1.

Each feature in classifying SSS from controls was compared with Receiver Operator Characteristics (ROC) analyses. Area Under the Curve (AUC) is used to determine which feature is best at classifying SSS from Controls. In Addition, we also use discrimination power (Equation 4).

$$DP = \frac{sqrt(3)}{\pi}(log(X) + log(Y)) \tag{4}$$

where: $X = sensitivity/(1sensitivity)$ and $Y = specificity/(1specificity)$

Optimal cutoff's are calculated using Youden's criterias.

The results from the ROC analysis for each features are summarized in Table 5,sorted by the highest AUC and Figure 2. The results suggest that best AUC with the permuted validity score (AUC = 80.09, cut-off = 0.17) and then the average normalized perimeter between the repetitions of each stimuli (AUC = 78.50, cut-off = 1.92 z-scores). Interestingly, while the permuted validity cut-off leads to higher specificity (77.50 vs. 74.29), the opposite is true for the sensitivity criteria (73.60 vs. 70.40).

For the concern that some criteria might bias towards types of synesthesia, we compared the groups by sub-type of SSS (i.e. weekdays, months and numbers) with the classifications using the cut-offs for permuted validity and perimeter. Figure 4 shows the venn diagram including only data from Ward (since we don't have the details from the other datasets). This suggests that 11% of the subsample in

Further analyses aimed at estimating the reliability of features when sub sampling the

dataset with different slices. First Section B we recalculated the ROC by sub sampling the data from the most extreme groups. The extreme groups were defined by the percentiles on the questionnaire sores, hence only the data from Ward is included there. We computed AUC, sensitivity and sensibility for the 10-90 %, 20-80 %, 30-70 %, 40-60 % subsampled participants depending on the dsistribution of the syneshtesia questionnaire (see Ward et al., 2018). The results for AUC Figure B1, sensitivity Figure B2 and specificity Figure B3.

However we also found differences across the datasets, see Figure B5 for AUC, Figure B6 for sensitivity and Figure B7 for specificity.

To confirm that the results are not circular, we correlated questionnaire scores with the features results, see Section B, Figure B4

Finally, descriptively we also wanted to see whether one of the main criteria would be more beneficial for one form of SSS or the other, see

Phase II Methods

Additional data using the same task will be collected in the future. The procedure will be the same as for the previously decribed task only that this time there will be four repetitions. We aim at extracting the same criteria on this new dataset and compare whether we can accurately predict the groups based on the thresholds described here.

Phase II Materials:

Materials are more details on the procedure are described here https://osf.io/pjb6e/?view_only=d467ebf4c1f94076ae4ac61298255065.

Phase II Planned population

https://osf.io/6h8dx

Discussion

Shifting from investigating consistency across stimulus position to across repetitions have led to some improvement in ROC. The best criteria was a GLM

From the different features we extracted, topological validity across the repetitions

appeared to be the one leading to the largest Area Under the Curve.

Limitations

Although an optimal tool to discriminate SSS might be particularly relevant for experimental purposes, it is important to consider some limitations. These consistency tools are designed with a limited set of sequential stimuli (i.e. months, weeks and the first ten natural numbers). Other sequences might also be represented in particular spatial positions such as temperature, ect. Another point is that rather than categorical, synesthesia might be present on a continuum in the general population. In that case diagnostic cutoffs might not be relevant, rather a score would be necessary (Price & Pearson, 2013). Finally, there might also be an issue of circularity - as with many diagnostics : how synesthesia is defined determines how synestetes are detected which are the groups on which synesthesia is defined (Simner, 2012). This is particularly relevant when the two diagnostic criteria on which validity are compared are self-reports (i.e. being conscious) and consistency.

The heterogeneity of methods used to detect SSS, combined with the heterogeneity of SSS complicates the task to estimate general population prevalence of SSS (Brang et al., 2010; Jonas & Price, 2014; Sagiv et al., 2006). For the present stake, the circularity issue makes it difficult since consistency diagnostic tools are designed to best classify control from synesthetes and hence depend on how those groups are initially defined. For example Section B suggests that the best AUC is given by different features depending on the different datasets.

Numerals. While weekdays and months are finite sets, numerals are infinite. It is possible that some criteria could improve wen taking a larger set of numerals into account, in particular since there many descriptively interesting form occur at different decimals (in hindo-arabic decimal number system), see examples in (Galton, 1880) .

SSS with 3D representations might also be under-diagnosed since the test is in 2D. However it seems that most SSS are relatively good at transposing 3D to 2D, which might be also explain by a more general advantage in visuo-spatial memory for SSS (Brang et al., 2010).

Overlapping responses. A methodological issue concerns participants that give the same

responses across conditions. These responses are a complication since we can't infer whether those conditions did not give rise to a synesthetic response in a synthesete or whether it is from a control that was confused about the instructions. On a methodological level, those responses can critically bias the diagnostic criteria. On one side excluding those responses would imbalance the number of responses by participant, on the other side including these responses might bias the diagnostic.

Future studies could use machine learning and/or neural network in an attempt to fin the best criteria for classifying synesthetes from control. This approach however need to have a clear explanability, since the main use of a criteria is experimental. Ideally, we would need an algorithm which could give individual probability to have SSS on which a threshold would help to

See also (Root et al., 2025).

Conclusions

The feature that led to the best AUC was topological validity. If confirmed, this might lead to interesting conclusions about how SSS map ordinal stimuli in space. The parallel between maps and neuroscience has a long history (i.e. retinotopy, sonotopy or somatotopy), hence it seems that the automatic spatial associations in SSS follow to some extend some topological rules (Eagleman, 2009).

The optimal criterion needs to be informed about the order between inducers (i.e. to construct the polygons) and interestingly suggests that synthetic inducer are structurally mapped following topological rules analogous to geographical space structures. Hence suggesting a spatial nature for the synthetic forms of space sequence synesthetes.

Interestingly, ordinality is very important an important semantic property of numbers (REF). Moreover that numbers are acquired sequentially (i.e. 1 is learned before 2) (REF). Hence the importance of ordinality in SSS is coherent with developmental accounts of Synesthesia (Price & Pearson, 2013).

References

Baron-Cohen, S., Harrison, J., Goldstein, L. H., & Wyke, M. (1993). Coloured Speech Perception: Is Synaesthesia what Happens when Modularity Breaks Down? *Perception*, *22*(4), 419–426. https://doi.org/10.1068/p220419

Brang, D., Teuscher, U., Ramachandran, V. S., & Coulson, S. (2010). Temporal sequences, synesthetic mappings, and cultural biases: The geography of time. *Consciousness and Cognition*, *19*(1), 311–320. https://doi.org/10.1016/j.concog.2010.01.003

Dixon, M. J., Smilek, D., & Merikle, P. M. (2004). Not all synaesthetes are created equal: Projector versus associator synaesthetes. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(3), 335–343. https://doi.org/10.3758/CABN.4.3.335

Eagleman, D. M. (2009). The objectification of overlearned sequences: A new view of spatial sequence synesthesia. *Cortex*, *45*(10), 1266–1277. https://doi.org/10.1016/j.cortex.2009.06.012

Galton, F. (1880). Visualised Numerals. *Nature*, *21*(533), 252–256. https://doi.org/10.1038/021252a0

Gould, C., Froese, T., Barrett, A. B., Ward, J., & Seth, A. K. (2014). An extended case study on the phenomenology of sequence-space synesthesia. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00433

Jonas, C. N., & Price, M. C. (2014). Not all synesthetes are alike: Spatial vs. Visual dimensions of sequence-space synesthesia. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01171

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Price, M., & Pearson, D. (2013). Toward a visuospatial developmental account of sequence-space synesthesia. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00689

Root, N., Asano, M., Melero, H., Kim, C.-Y., Sidoroff-Dorso, A. V., Vatakis, A., Yokosawa, K., Ramachandran, V., & Rouw, R. (2021). Do the colors of your letters depend on your

language? Language-dependent and universal influences on grapheme-color synesthesia in seven languages. *Consciousness and Cognition*, *95*, 103192. https://doi.org/10.1016/j.concog.2021.103192

Root, N., Chkhaidze, A., Melero, H., Sidoroff-Dorso, A., Volberg, G., Zhang, Y., & Rouw, R. (2025). How "diagnostic" criteria interact to shape synesthetic behavior: The role of self-report and test–retest consistency in synesthesia research. *Consciousness and Cognition*, *129*, 103819. https://doi.org/10.1016/j.concog.2025.103819

Rothen, N., Jünemann, K., Mealor, A. D., Burckhardt, V., & Ward, J. (2016a). The sensitivity and specificity of a diagnostic test of sequence-space synesthesia. *Behavior Research Methods*, *48*(4), 1476–1481. https://doi.org/10.3758/s13428-015-0656-2

Rothen, N., Jünemann, K., Mealor, A. D., Burckhardt, V., & Ward, J. (2016b). The sensitivity and specificity of a diagnostic test of sequence-space synesthesia. *Behavior Research Methods*, *48*(4), 1476–1481. https://doi.org/10.3758/s13428-015-0656-2

Rothen, N., Seth, A. K., Witzel, C., & Ward, J. (2013). Diagnosing synaesthesia with online colour pickers: Maximising sensitivity and specificity. *Journal of Neuroscience Methods*, *215*(1), 156–160. https://doi.org/10.1016/j.jneumeth.2013.02.009

Sagiv, N., Simner, J., Collins, J., Butterworth, B., & Ward, J. (2006). What is the relationship between synaesthesia and visuo-spatial number forms? *Cognition*, *101*(1), 114–128. https://doi.org/10.1016/j.cognition.2005.09.004

Seron, X., Pesenti, M., Noël, M.-P., Deloche, G., & Cornet, J.-A. (1992). Images of numbers, or "when 98 is upper left and 6 sky blue". *Cognition*, *44*(1), 159–196. https://doi.org/10.1016/0010-0277(92)90053-K

Simner, J. (2012). Defining synaesthesia. *British Journal of Psychology*, *103*(1), 1–15. https://doi.org/10.1348/000712610X528305

Smilek, D., Callejas, A., Dixon, M. J., & Merikle, P. M. (2007). Ovals of time: Time-space associations in synaesthesia. *Consciousness and Cognition*, *16*(2), 507–519. https://doi.org/10.1016/j.concog.2006.06.013

Van Petersen, E., Altgassen, M., Van Lier, R., & Van Leeuwen, T. M. (2020b). Enhanced spatial navigation skills in sequence-space synesthetes. *Cortex*, *130*, 49–63. https://doi.org/10.1016/j.cortex.2020.04.034

Van Petersen, E., Altgassen, M., Van Lier, R., & Van Leeuwen, T. M. (2020a). Enhanced spatial navigation skills in sequence-space synesthetes. *Cortex*, *130*, 49–63. https://doi.org/10.1016/j.cortex.2020.04.034

Ward, J. (n.d.-a). *Optimizing a Measure of Consistency for Sequence-Space Synaesthesia*. https://doi.org/10.31234/osf.io/5cnr7

Ward, J. (n.d.-b). *Optimizing a Measure of Consistency for Sequence-Space Synaesthesia*. https://doi.org/10.31234/osf.io/5cnr7

Ward, J., Ipser, A., Phanvanova, E., Brown, P., Bunte, I., & Simner, J. (2018). The prevalence and cognitive profile of sequence-space synaesthesia. *Consciousness and Cognition*, *61*, 79–93. https://doi.org/10.1016/j.concog.2018.03.012

Table 1

My Caption

Source	Synesthetes				Controls			
	Original N	Age	n females	Included N	N	Age	n females	Included
(Rothen et al., 2016a)	33	23.1	24	37	37	28.2	27	37
(Van Petersen et al., 2020b)	23	23.22	20	21	21	21.57	19	13
(Ward, n.d.-b)	252	37.21	202	249	215	19.90	178	204
Ward 2				88				17
Merged				395				271

Note. Note below table

Table 2

Summary of data sources

dataSource	Ctl	Syn
PeterCor	21	22
Rothen	37	32
Ward	213	252
Ward2	18	90

Note. Sources are

Table 3

Descriptives of each features

Table 4

Feature	Ctl	Syn
QuestScoreRL	45.01 (11.04)	22.65 (6.95)
Area_perc	0.47 (0.88)	0.26 (0.54)
Area_zs	0.27 (0.33)	0.08 (0.14)
Perimeter_zs	3.33 (1.72)	1.61 (1.17)
perm_zs	99.79 (790.15)	14.27 (217.45)
SelfInter	8.83 (10.45)	1.71 (4.69)
areaPoly_GA	1.03 (0.99)	1.86 (1.24)
perim_GA	9.61 (3.41)	8.39 (2.49)
isSimple_GA	0.21 (0.22)	0.38 (0.26)
isValidStruct_M	0.13 (0.19)	0.37 (0.26)
Corr_XY_M	0.08 (0.44)	-0.02 (0.29)
Corr_M	0.22 (0.28)	0.26 (0.17)
isValid_perm_M	0.12 (0.16)	0.35 (0.23)
GLM_Valid_areazs	-0.37 (1.13)	1.03 (1.25)

Table 5

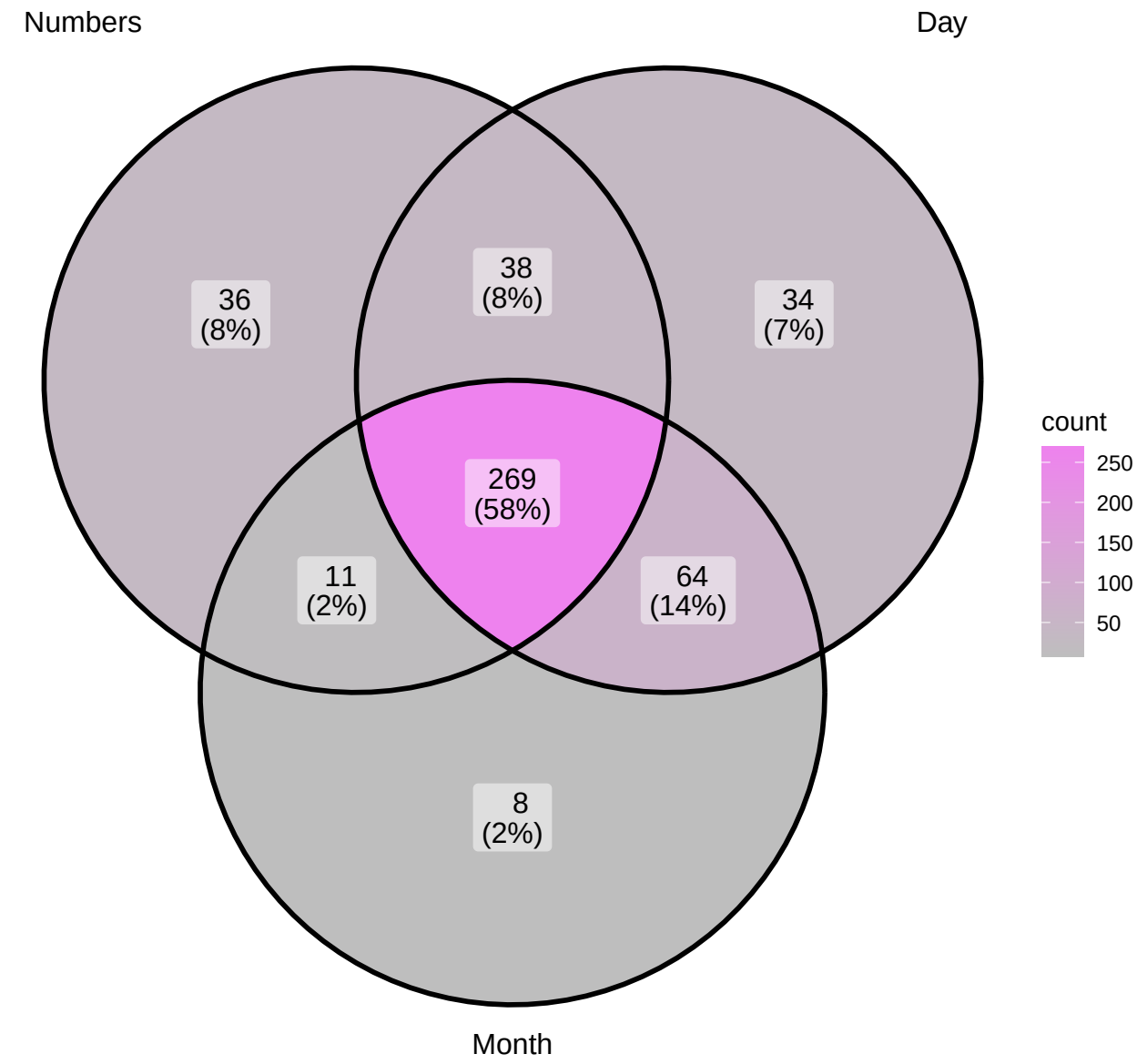
Summary of ROC analysis

Table 6

Feature	AUC	DP	threshold	sensitivity	specificity	ci_low	ci_high	power
GLM_Valid_areazs	80.33	2.12	-0.07	79.73	68.93	76.94	83.72	1.00
isValid_perm_M	80.09	2.06	0.17	70.40	77.50	76.69	83.50	1.00
Perimeter_zs	78.50	2.04	1.92	73.60	74.29	74.80	82.20	1.00
isValidStruct_M	77.07	1.81	0.17	71.73	71.43	73.53	80.61	1.00
SelfInter	71.71	1.68	1.17	79.47	58.93	67.60	75.82	1.00
areaPoly_GA	70.96	1.30	1.29	64.53	67.50	67.02	74.90	1.00
Area_zs	70.94	1.72	0.08	75.47	65.36	66.66	75.22	1.00
isSimple_GA	69.91	1.26	0.28	60.53	70.36	65.93	73.89	1.00
perm_zs	65.80	2.18	-3.67	93.87	37.86	61.40	70.21	1.00
perim_GA	60.65	1.32	10.46	83.47	43.21	56.03	65.27	1.00
Corr_XY_M	59.00	0.98	0.20	82.40	36.79	54.43	63.57	0.98
Corr_M	56.32	1.07	0.08	88.00	28.93	51.73	60.91	0.81
Area_perc	51.93	0.82	0.21	76.27	41.79	47.13	56.73	0.14

Figure 1

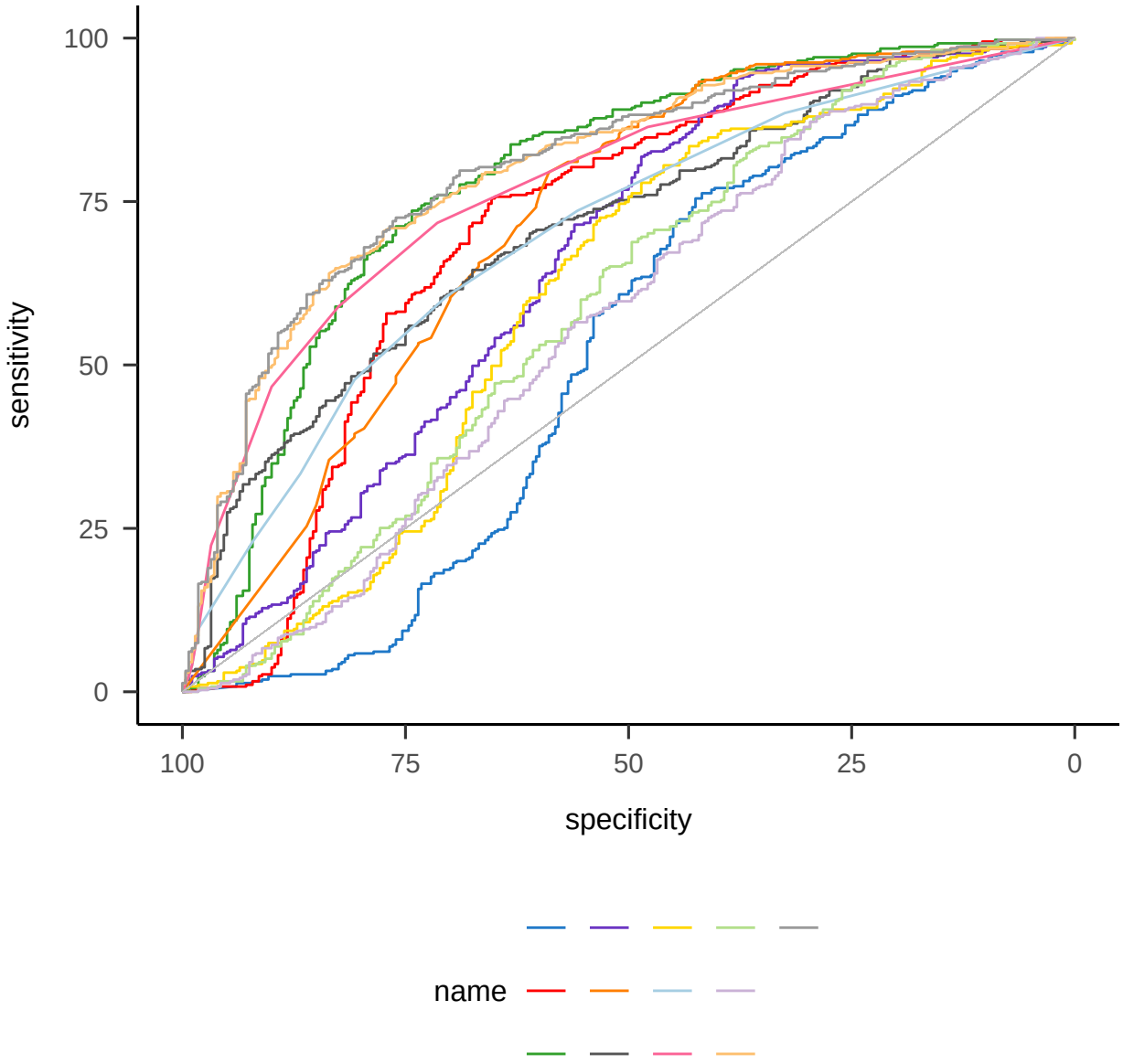
Venn diagram of the types of self-reported SSS



Note. Only the subsample from Ward is included here

Figure 2

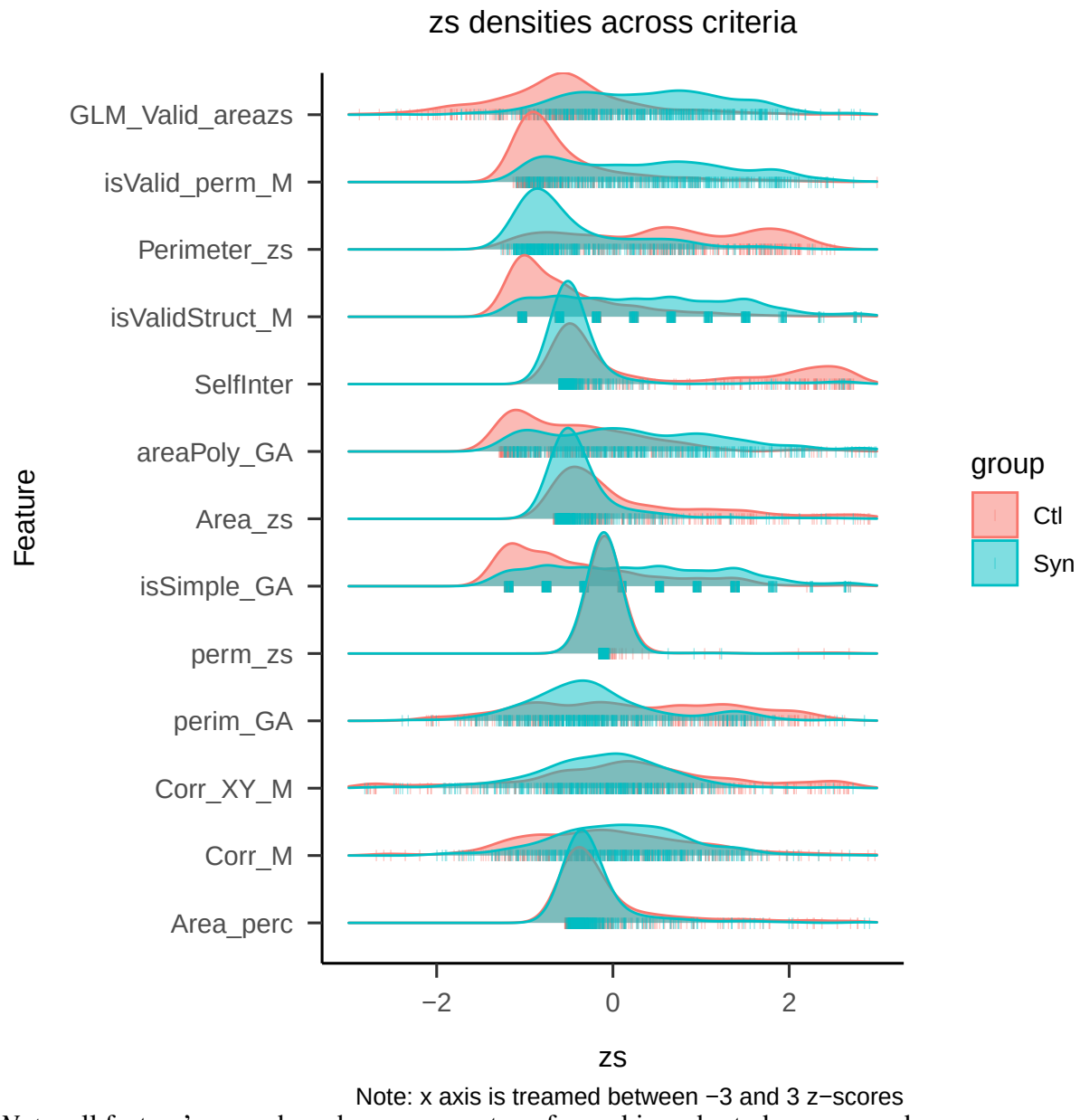
All the ROC curves for each features



Note. ROC curves for each features

Figure 3

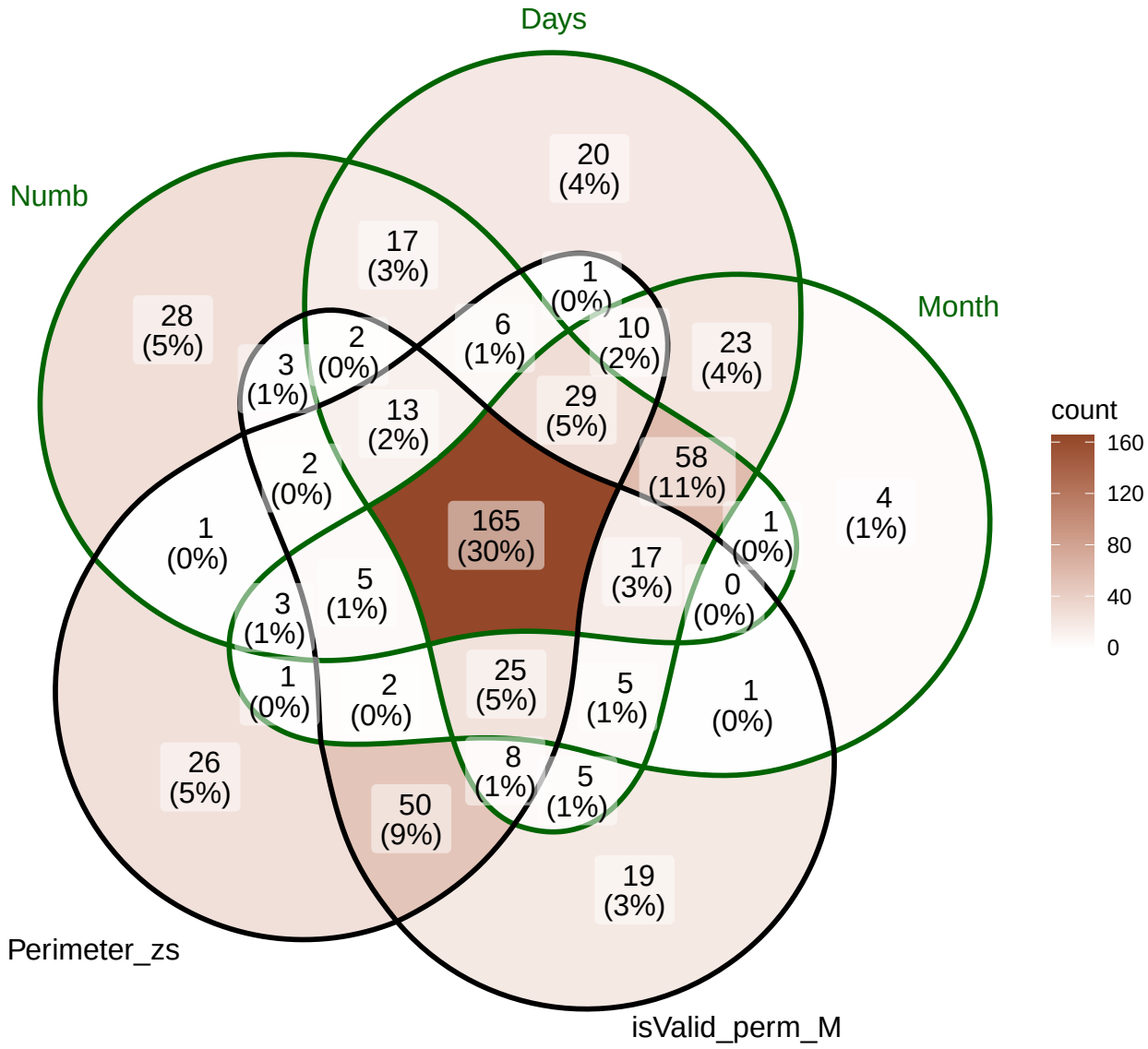
Density plots of all the features comparing SSS and controls



Note. all feature's score have been z-score transformed in order to be compared

Figure 4

Avocado Venn Diagramm of self-report and tests



Note. Only data from Ward is included here

Appendix A

Appendix B

To additionally test the validity of the criteria, we computed the ROC again by subsampling the groups based on the questionnaire scores so to have more extreme groups. This was done only on the data from Ward, since the other did not include a questionnaire in the data.

Appendix 1 Subsambled data by questionnaire quantiles (10% steps)

In the following, we compare the data sampled by the questionnaire score. Based on the distribution of the questionnaire score, we sampled the 10 % with the lowest and 10 % with the highest scores. Those are then compared with the 20 and 20 % and so on until 40 and 40 %. The rationale of this procedure is that AUC, sensitivity and specificity should remain stable across percentiles for a feature to be valid, see Figure B1. In other words the ROC should remain unchanged if we take extreme groups compared to less extreme ones.

Appendix 2 Correlation with self-report

The best criterion should also best correlate with SSS self-reported questionnaire score.

Works only with Ward's aggregated data, see Figure B2.

Appendix 3 By dataset

Here we compare the ROC for each specific data sample. Although the different authors have used a similar method, there might be a recruitment bias or other.

Table B1

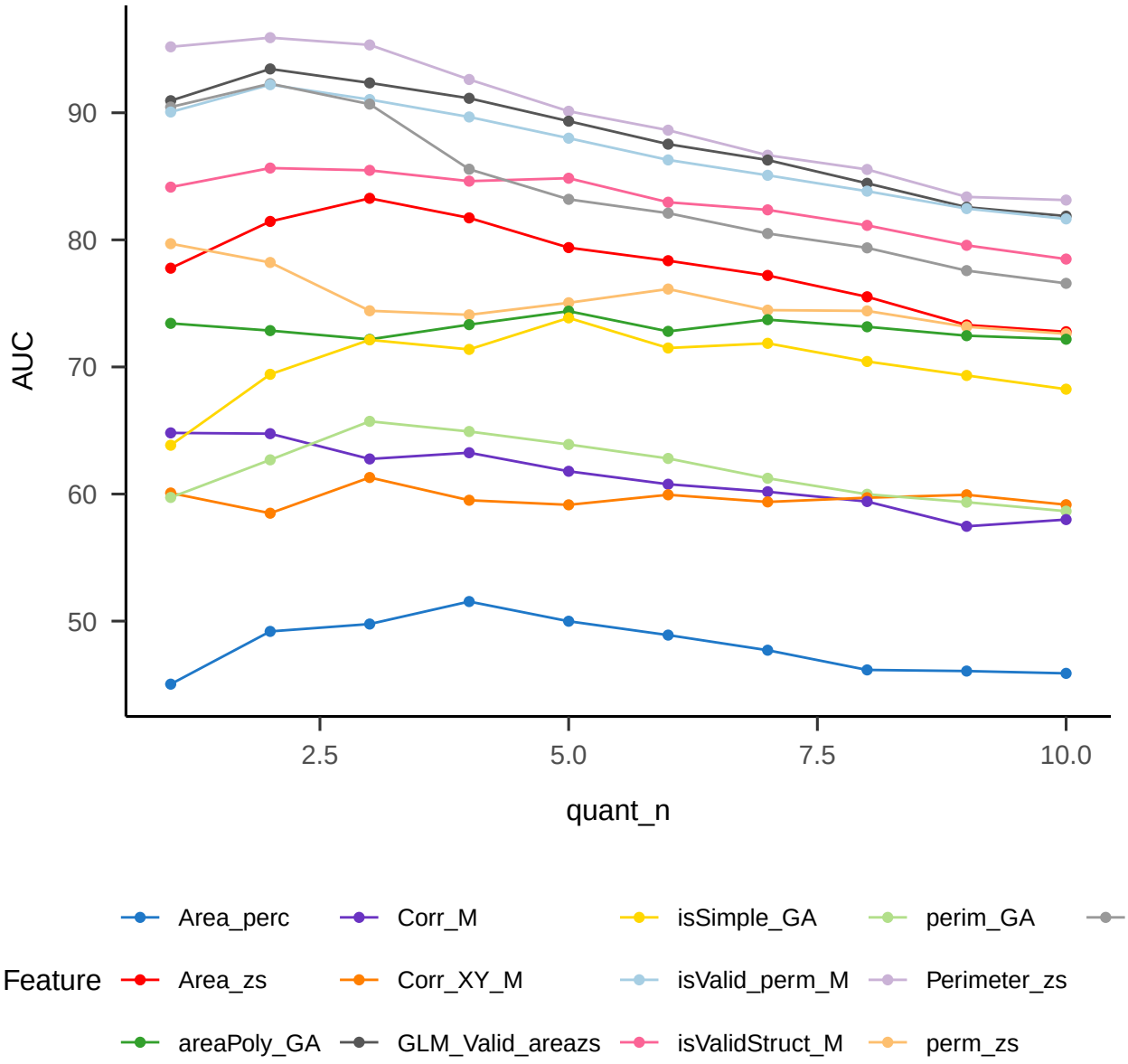
Figure B5

Figure B6

Figure B7

Figure B1

Lineplots of AUC by percentile



Note. Each point represents an increasing percentiles

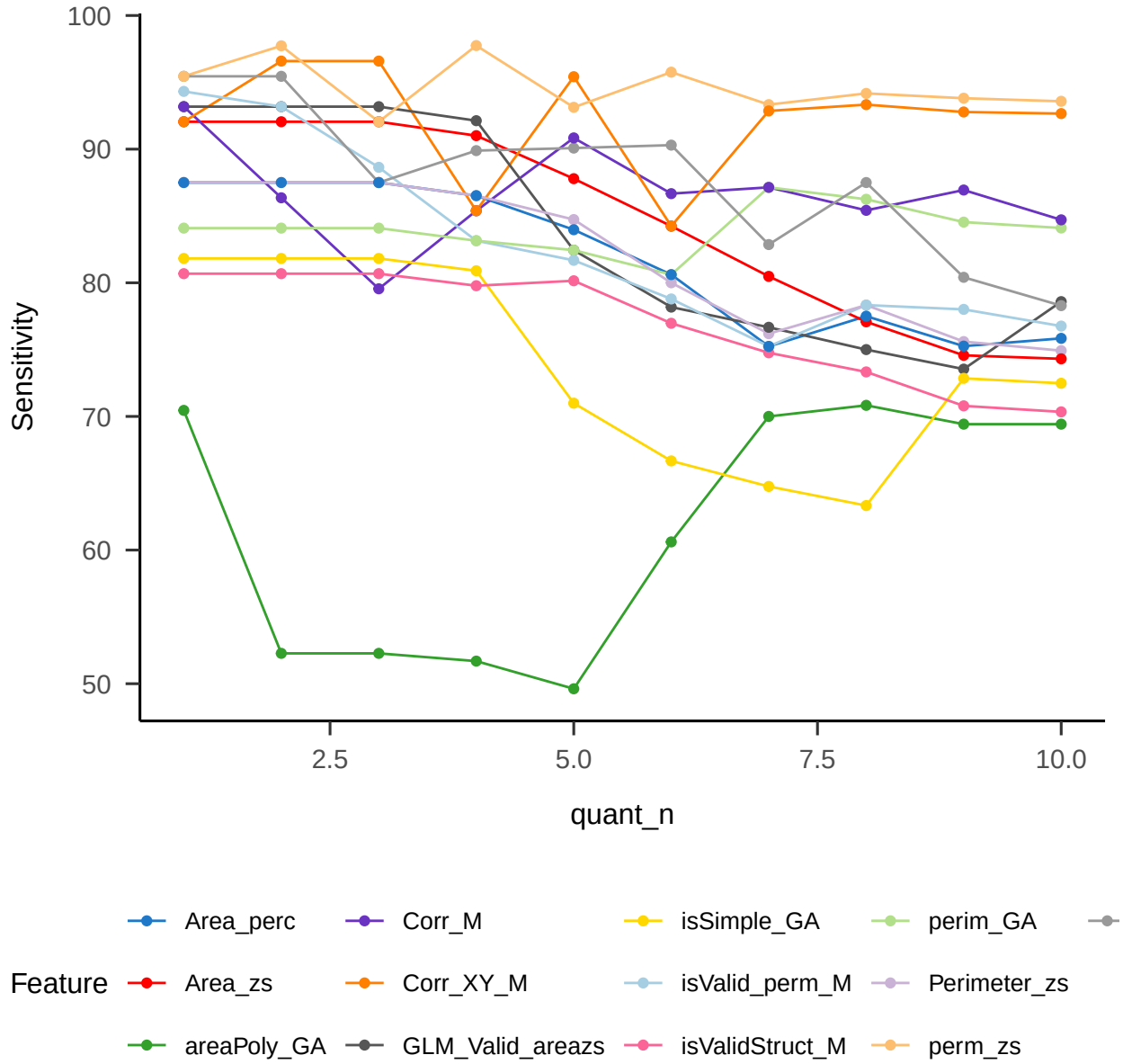
Appendix 4 Reliability

Appendix 5 Plot all

This exports many pdf's. It plots each ID and condition z-score x and y coordinates. Since each coordinate is repeated 3 times, these are represented by triangles. The line paths connect

Figure B2

Lineplots of Sensitivity by percentile

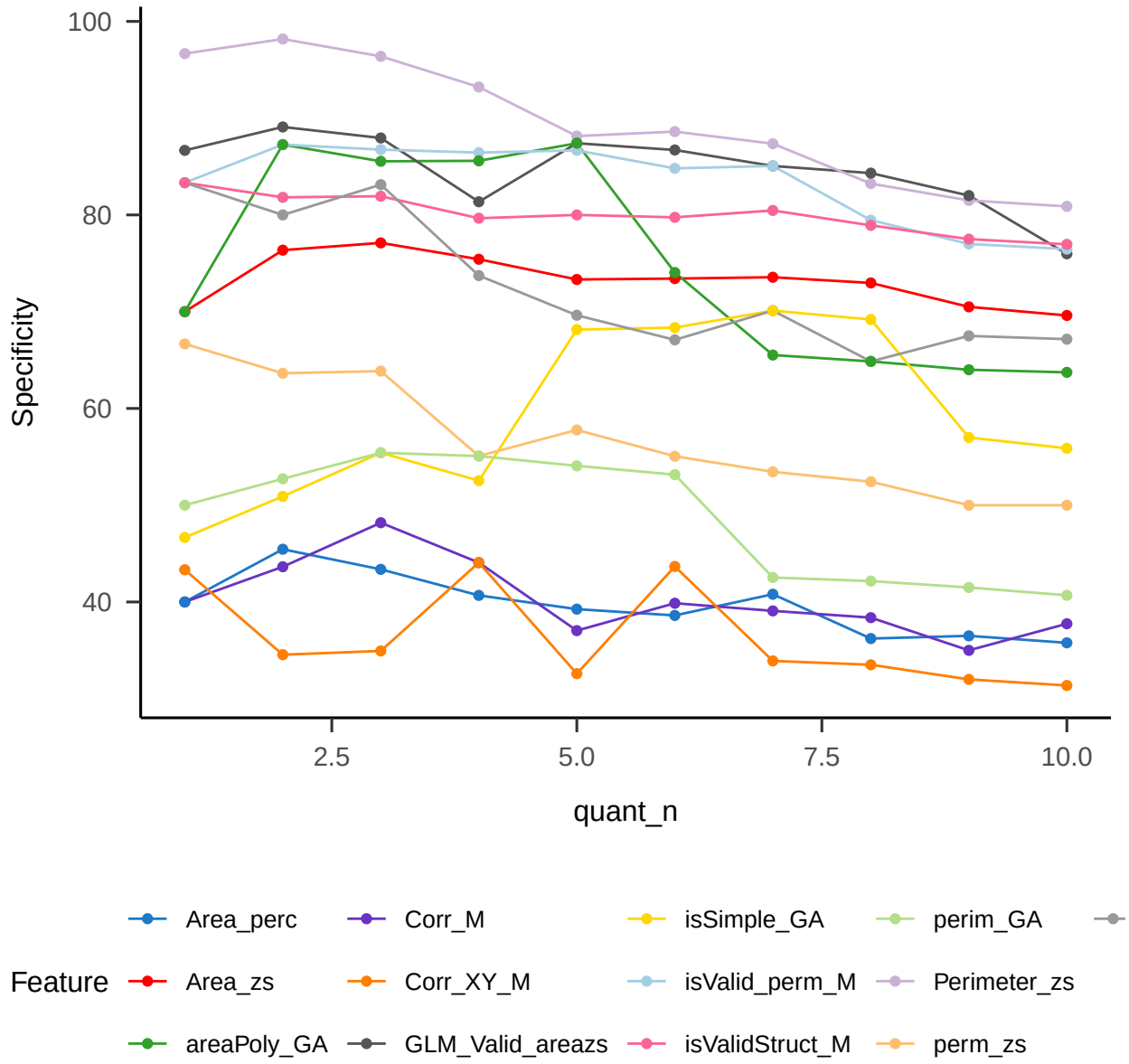


Note. Each point represents an increasing percentiles

average coordinates to visualize forms (stimulus are ordered, i.e. 1 to 9, Monday to Sunday, January to December). Finally in the top right corner, each dots indicates if the ID would pass / fails depending on the criteria.

Figure B3

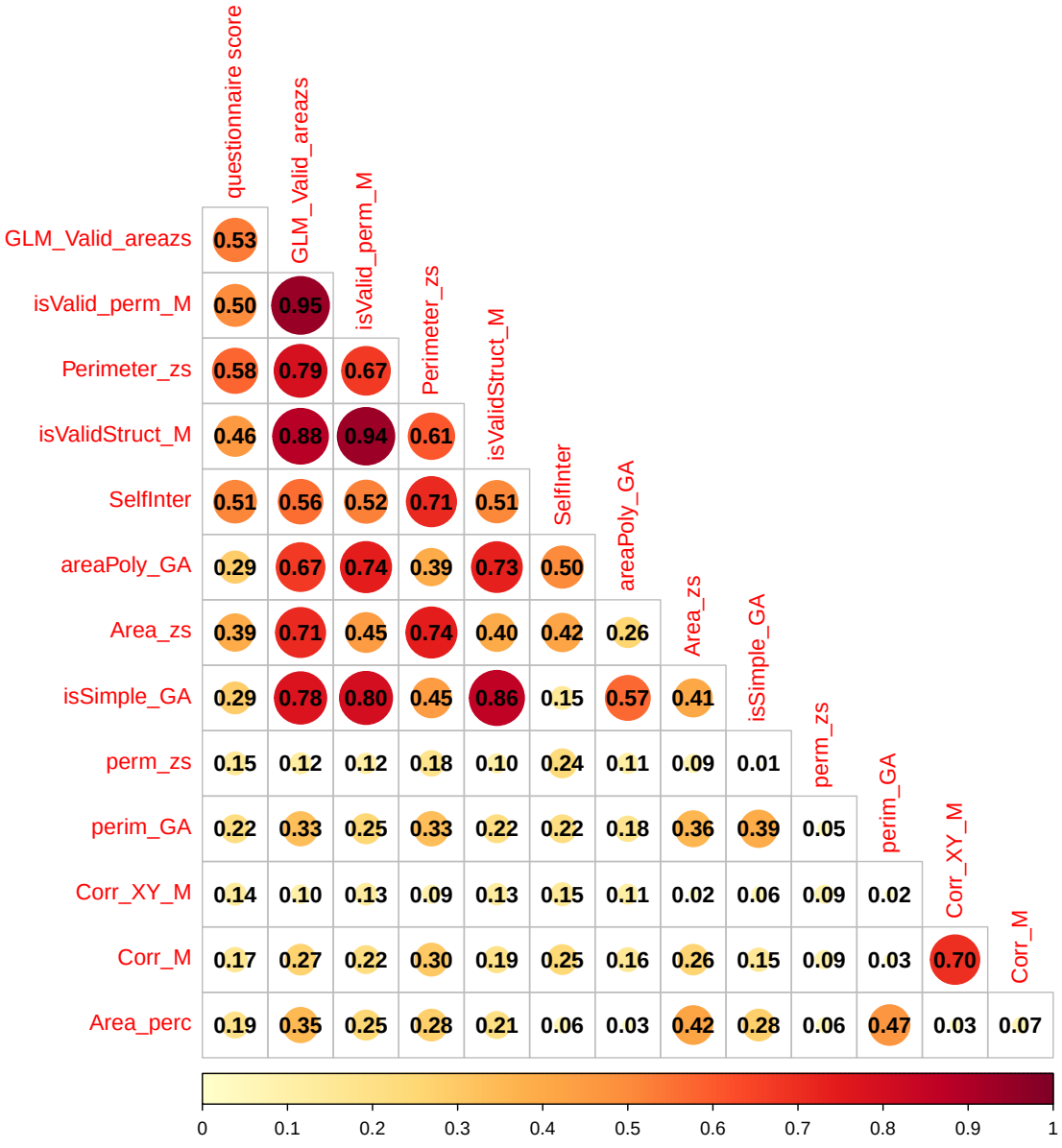
Lineplots of Specificity by percentile



Note. Each point represents an increasing percentiles

Figure B4

Correlation with self-reported questionnaire



Note. Only data from Ward is included here

Table B1

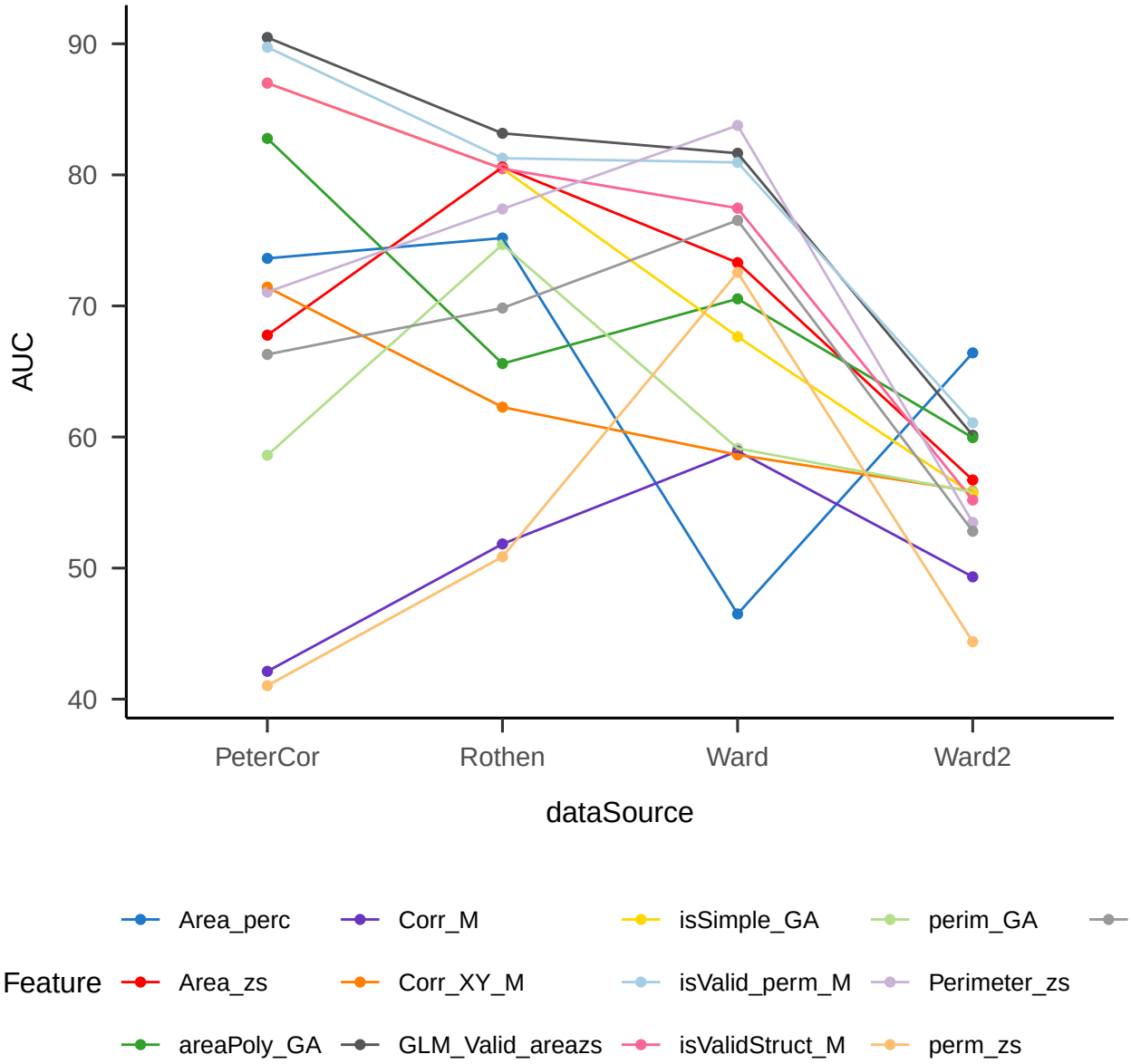
Average feature for each group and dataset

Table B2

dataSource	Feature	Ctl	Syn
PeterCor	QuestScoreRL	NaN (* NA*)	NaN (* NA*)
Rothen	QuestScoreRL	NaN (* NA*)	NaN (* NA*)
Ward	QuestScoreRL	45.01 (*11.04*)	22.65 (*6.95*)
Ward2	QuestScoreRL	NaN (* NA*)	NaN (* NA*)
PeterCor	Area_perc	0.31 (*0.40*)	0.07 (*0.05*)
Rothen	Area_perc	0.89 (*1.44*)	0.13 (*0.09*)
Ward	Area_perc	0.38 (*0.69*)	0.26 (*0.50*)
Ward2	Area_perc	0.80 (*1.36*)	0.34 (*0.69*)
PeterCor	Area_zs	0.06 (*0.11*)	0.02 (*0.01*)
Rothen	Area_zs	0.23 (*0.25*)	0.05 (*0.06*)
Ward	Area_zs	0.30 (*0.34*)	0.08 (*0.15*)
Ward2	Area_zs	0.19 (*0.33*)	0.09 (*0.15*)
PeterCor	Perimeter_zs	1.33 (*0.79*)	0.77 (*0.24*)
Rothen	Perimeter_zs	2.91 (*1.37*)	1.60 (*0.96*)
Ward	Perimeter_zs	3.71 (*1.65*)	1.61 (*1.19*)
Ward2	Perimeter_zs	2.25 (*1.85*)	1.75 (*1.21*)
PeterCor	perm_zs	-5.90 (*0.72*)	-5.82 (*0.57*)
Rothen	perm_zs	-5.63 (*1.28*)	-5.78 (*0.79*)
Ward	perm_zs	139.11 (*923.23*)	12.42 (*223.34*)
Ward2	perm_zs	-5.83 (*0.98*)	27.03 (*242.33*)
PeterCor	SelfInter	1.41 (*2.22*)	0.35 (*0.46*)
Rothen	SelfInter	1.78 (*2.37*)	0.36 (*0.47*)
Ward	SelfInter	11.39 (*11.05*)	1.84 (*4.92*)
Ward2	SelfInter	3.02 (*5.34*)	1.85 (*4.87*)
PeterCor	areaPoly_GA	0.63 (*0.68*)	1.99 (*1.24*)

Figure B5

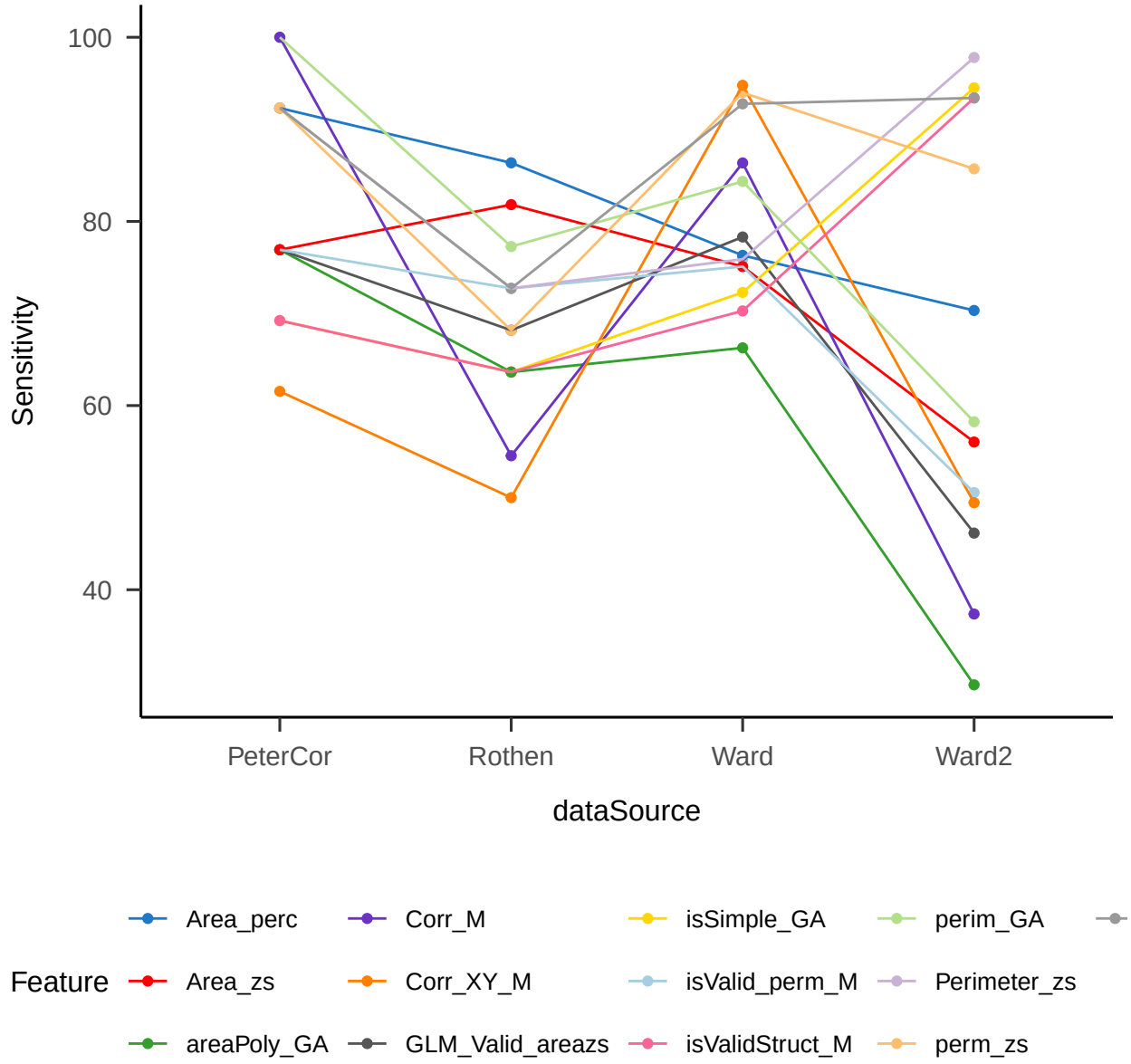
Lineplots of AUC by data source



Note. AUC

Figure B6

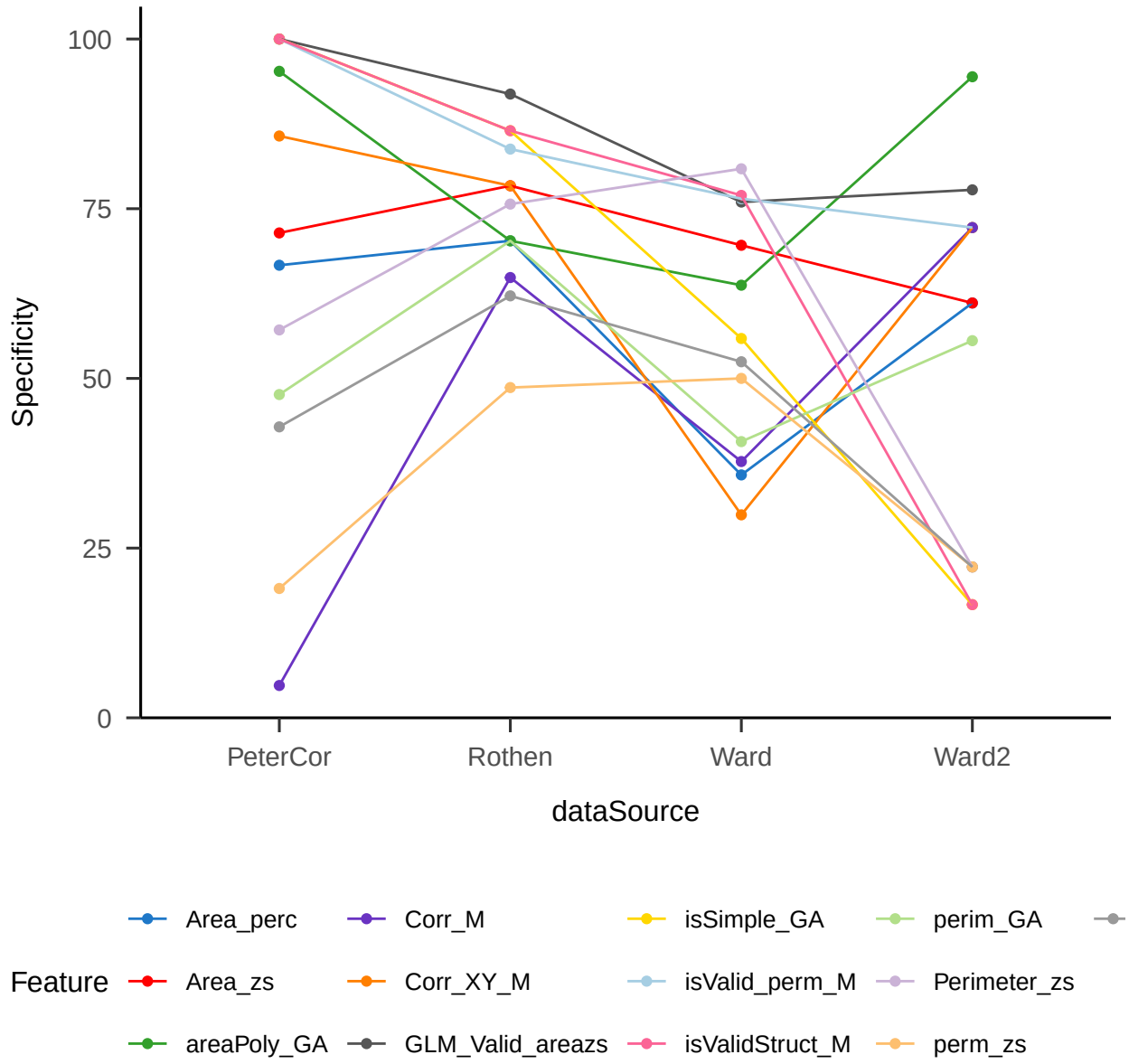
Lineplots of Sensitivity by data source



Note. Sensitivity

Figure B7

Lineplots of Specificity by data source



Note. Specificity

	Feature	AUC	DP	threshold	sensitivity	specificity	ci_low	ci_high	power
12	isValid_perm_M	81.72	2.30	0.22	68.51	82.88	77.15	86.29	1.00
13	GLM_Valid_areas	81.25	2.27	0.15	76.24	76.03	76.60	85.89	1.00
3	Perimeter_zs	79.13	2.12	1.93	73.48	76.03	74.08	84.17	1.00
9	isValidStruct_M	77.38	1.87	0.17	71.27	73.29	72.42	82.34	1.00
2	Area_zs	71.82	1.81	0.08	77.35	65.07	65.99	77.65	1.00
6	areaPoly_GA	71.46	1.46	1.11	71.82	63.70	65.91	77.01	1.00
5	SelfInter	70.84	1.75	1.67	83.98	53.42	65.02	76.65	1.00
8	isSimple_GA	70.81	1.33	0.28	61.88	70.55	65.28	76.34	1.00
4	perm_zs	66.95	1.59	-5.09	82.87	51.37	60.77	73.12	1.00
7	perim_GA	59.41	1.26	10.44	82.32	43.84	52.91	65.91	0.85
10	Corr_XY_M	59.23	0.88	0.07	65.75	56.16	52.82	65.64	0.84
11	Corr_M	56.85	1.02	0.10	85.64	32.19	50.45	63.26	0.59
1	Area_perc	55.69	1.16	0.19	78.45	47.26	49.00	62.38	0.44

	Feature	AUC	DP	threshold	sensitivity	specificity	ci_low	ci_high	power
13	GLM_Valid_areazs	79.65	2.15	-0.11	79.90	69.40	74.66	84.64	1.00
12	isValid_perm_M	78.48	1.95	0.17	68.04	77.61	73.39	83.58	1.00
3	Perimeter_zs	77.68	1.97	2.02	76.80	69.40	72.20	83.17	1.00
9	isValidStruct_M	76.72	1.73	0.17	72.16	69.40	71.63	81.81	1.00
5	SelfInter	72.84	1.76	1.17	80.41	59.70	67.05	78.64	1.00
6	areaPoly_GA	70.48	1.49	1.75	50.00	82.09	64.85	76.11	1.00
2	Area_zs	69.97	1.67	0.07	74.23	65.67	63.64	76.30	1.00
8	isSimple_GA	68.91	1.20	0.28	59.28	70.15	63.14	74.68	1.00
4	perm_zs	64.62	2.27	-3.78	94.33	38.06	58.28	70.96	1.00
7	perim_GA	61.89	1.35	9.75	80.41	49.25	55.30	68.49	0.97
10	Corr_XY_M	58.66	1.12	0.20	84.54	36.57	52.10	65.21	0.78
11	Corr_M	55.90	1.17	0.07	88.66	29.85	49.26	62.53	0.46
1	Area_perc	48.24	1.05	0.63	91.75	20.90	41.30	55.17	0.08