

# 000 001 002 003 004 005 CONFORMAL CONFIDENCE SETS FOR BIOMEDICAL 006 IMAGE SEGMENTATION 007 008 009

010 **Anonymous authors**  
 011 Paper under double-blind review  
 012  
 013  
 014  
 015  
 016  
 017  
 018  
 019  
 020  
 021  
 022  
 023  
 024

## ABSTRACT

011 We develop confidence sets which provide spatial uncertainty guarantees for the  
 012 output of a black-box machine learning model designed for image segmentation.  
 013 To do so we adapt conformal inference to the imaging setting, learning thresholds  
 014 on a calibration dataset based on the distribution of the maximum of the trans-  
 015 formed logit scores within and outside of the ground truth masks. We show that  
 016 these confidence sets, when applied to new predictions of the model, are guaran-  
 017 teed to contain the true unknown segmented mask with desired probability. We  
 018 illustrate and validate our approach on a polyps tumor segmentation dataset. To  
 019 do so we obtain the logit scores from a deep neural network trained for polyps  
 020 segmentation and show that using distance transformed scores to obtain outer con-  
 021 fidence sets and the original scores for inner confidence set enables tight bounds  
 022 on tumor location whilst controlling the false coverage rate.  
 023  
 024

## 1 INTRODUCTION

027 Deep neural networks promise to significantly enhance a wide range of important tasks in biomedical  
 028 imaging. However these models, as typically used, lack formal uncertainty guarantees on their out-  
 029 put which can lead to overconfident predictions and critical errors. Misclassifications or inaccurate  
 030 segmentations can lead to serious consequences, including misdiagnosis, inappropriate treatment  
 031 decisions, or missed opportunities for early intervention. As a consequence, despite their potential  
 032 utility, medical professionals cannot yet rely on deep learning models to provide accurate informa-  
 033 tion and predictions which greatly limits their use in practical applications.

034 In order to address this problem, conformal inference, a robust framework for uncertainty quan-  
 035 tification, has become increasingly used as a means of providing prediction guarantees, offering  
 036 reliable, distribution-free confidence sets for the output of neural networks which have finite sample  
 037 validity. This approach, originally introduced in XXX, has become increasingly popular (CITE)  
 038 due to its ability to provide rigorous statistical guarantees without making strong assumptions about  
 039 the underlying data distribution or model architecture. Conformal prediction methods, in their most  
 040 commonly used form - split conformal inference - work by calibrating the predictions of the model  
 041 on a held-out dataset in order to provide sets which contain the output with a given probability, see  
 042 Angelopoulos & Bates (2021) for a good introduction.

043 In the context of image segmentation, we have a decision to make at each pixel/voxel of an im-  
 044 age which can lead to a large multiple testing problem. Traditional conformal methods, typically  
 045 designed for scalar outputs, require adaptation to handle multiple tests and their inherent spatial  
 046 dependencies. To do so Angelopoulos et al. (2021) applied conformal inference pixelwise and per-  
 047 formed multiple testing correction on the resulting  $p$ -values, however this approach does not take  
 048 into account of the complex dependence structure inherent in the images. To take advantage of this  
 049 structure, in an approach analogous to the FDR control of (Benjamini & Hochberg, 1995), Bates  
 050 et al. (2021) and Angelopoulos et al. (2022) sought to control the expected risk of a given loss func-  
 051 tion over the image and used a conformal approach to produce outer confidence sets for segmented  
 052 images which control the expected false negative rate. Other work considering conformal inference  
 053 in the context of multiple dependent hypotheses include XXX and XXX who established conformal  
 FDR control when testing for the presence of missing links in graphs. Under exchangeability  
 of the considered hypotheses XXX provides false coverage rate control over multiple conformal

054 inferences. XXX and XXX instead applied conformal inference in the imaging setting but did not  
 055 account for multiple comparisons.  
 056

057 In this work we argue that bounding the segmented outcome with guarantees in probability rather  
 058 than in expectation/proportion can be more informative, avoiding errors at the borders of potential  
 059 tumors. This is analogous to the tradeoff between FWER and FDR/FDP control in the multiple testing  
 060 literature in which there is a balance between power and coverage rate, the distinction being that  
 061 in medical image segmentation there can be a potentially serious consequence to making mistakes.  
 062 Under-segmentation might cause part of the tumor to be missed, potentially leading to inadequate  
 063 treatment. Over-segmentation, on the other hand, could result in unnecessary interventions, increasing  
 064 patient risk and healthcare costs. Unlike bounds on the proportion of discovered pixels/voxels,  
 065 confidence sets are guaranteed to contain the outcome with a given level of confidence and allow  
 066 medical practitioners to follow-up on the images where there is greater uncertainty. Since the guar-  
 067 antees are more meaningful the problem is more difficult. As we shall see, using the original scores  
 068 can lead to rather large and uninformative outer confidence sets. In order to address this, we use  
 069 a held out learning dataset to learn the score transformations which provide the most informative  
 070 confidence regions.

071 In order to obtain confidence sets we use a split-conformal inference approach in which we learn  
 072 appropriate cutoffs, with which to threshold the output of an image segmenter, from a calibration  
 073 dataset. These thresholds are obtained by considering the distribution of the maximum logit (trans-  
 074 formed) scores provided by the model within and outside of the ground truth masks. This approach  
 075 allows us to capture the spatial nature of the uncertainty in segmentation tasks, going beyond simple  
 076 pixel-wise confidence measures. By applying these learned thresholds to new predictions, we can  
 077 generate confidence sets that are guaranteed to contain the true, unknown segmented mask with a  
 078 desired probability.

## 079 2 THEORY

### 080 2.1 SET UP

082 Let  $\mathcal{V} \subset \mathbb{R}^m$ , for some dimension  $m \in \mathbb{N}$ , be a finite set corresponding to the domain which  
 083 represents the pixels/voxels at which we observe imaging data. Let  $\mathcal{X} = \{g : \mathcal{V} \rightarrow \mathbb{R}\}$  be the set  
 084 of real functions on  $\mathcal{V}$  and let  $\mathcal{Y} = \{g : \mathcal{V} \rightarrow \{0, 1\}\}$  be the set of all functions taking the values  
 085 0 or 1. We shall refer to elements of  $\mathcal{X}$  and  $\mathcal{Y}$  as images. Suppose that we observe a calibration  
 086 dataset  $(X_i, Y_i)_{i=1}^n$  of random images, where  $X_i : \mathcal{V} \rightarrow \mathbb{R}$  represents the  $i$ th observed calibration  
 087 image and  $Y_i : \mathcal{V} \rightarrow \{0, 1\}$  outputs labels at each  $v \in \mathcal{V}$  giving 1s at the true location of the objects  
 088 in the image  $X_i$  that we wish to identify and 0s elsewhere. Let  $\mathcal{P}(\mathcal{V})$  be the set of all subsets of  $\mathcal{V}$ .  
 089 Moreover, given a function  $f : \mathcal{X} \rightarrow \mathcal{X}$ , we shall write  $f(X, v)$  to denote  $f(X)(v)$  for all  $v \in \mathcal{V}$ .

090 Let  $s : \mathcal{X} \rightarrow \mathcal{X}$  be a score function - trained on an independent dataset - such that given an image  
 091 pair  $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ ,  $s(X)$  is a score image in which  $s(X, v)$  is intended to be higher at the  $v \in \mathcal{V}$   
 092 for which  $Y(v) = 1$ . The score function can for instance be the logit scores obtained from a deep  
 093 neural network image segmentation method to the image  $X$  as input e.g. CITE. Given  $X \in \mathcal{X}$ , let  
 094  $\hat{M}(X) \in \mathcal{Y}$  be the predicted mask based on the original segmentation model.

095 In what follows we will use the calibration dataset to construct a confidence functions  $I, O : \mathcal{X} \rightarrow$   
 096  $\mathcal{P}(\mathcal{V})$  such that for a new image pair  $(X, Y) \sim \mathcal{D}$ , given error rates  $\alpha_1, \alpha_2 \in (0, 1)$  we have

$$\mathbb{P}(I(X) \subseteq \{v \in \mathcal{V} : Y(v) = 1\}) \geq 1 - \alpha_1, \quad (1)$$

$$\text{and } \mathbb{P}(\{v \in \mathcal{V} : Y(v) = 1\} \subseteq O(X)) \geq 1 - \alpha_2. \quad (2)$$

100 Here  $I(X)$  and  $O(X)$  serve as inner and outer confidence sets for the location of the true segmented  
 101 mask. Their interpretation is that, up to the guarantees provided by the probabilistic statements (1)  
 102 and (9), we can be sure that for each  $v \in I(X)$ ,  $Y(v) = 1$  or that for each  $v \notin O(X)$ ,  $Y(v) = 0$ .  
 103 See Figure 3 for an example of this in practice. Joint control over the events can also be guaranteed,  
 104 either by sensible choices of  $\alpha_1$  and  $\alpha_2$  or by using the joint distribution of the maxima of the logit  
 105 scores - see Section 2.3.

106 In order to establish conformal confidence results we shall require the following exchangeability  
 107 assumption.

108     **Assumption 1.** Given a new random image pair,  $(X_{n+1}, Y_{n+1})$ , suppose that  $(X_i, Y_i)_{i=1}^{n+1}$  is an  
 109     exchangeable sequence of random image pairs in the sense that  
 110

$$111 \quad \{(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})\} =_d \{(X_{\sigma(1)}, Y_{\sigma(1)}), \dots, (X_{\sigma(n+1)}, Y_{\sigma(n+1)})\}$$

112     for any permutation  $\sigma \in S_{n+1}$ . Here  $=_d$  denotes equality in distribution and  $S_{n+1}$  is the group of  
 113     permutations of the integers  $\{1, \dots, n+1\}$ .  
 114

115     Exchangeability or a variant is a standard assumption in the conformal inference literature (An-  
 116     gelopoulos & Bates, 2021) and facilitates coverage guarantees. It holds for instance if we assume  
 117     that the collection  $(X_i, Y_i)_{i=1}^{n+1}$  is an i.i.d. sequence of image pairs but is more general and in prin-  
 118     ciple allows for other dependence structures.  
 119

## 120     2.2 MARGINAL CONFIDENCE SETS

121     In order to construct conformal confidence sets let  $f_I, f_O : \mathcal{X} \rightarrow \mathcal{X}$  be inner and outer trans-  
 122     formation functions and for each  $1 \leq i \leq n+1$ , let  $\tau_i = \max_{v \in \mathcal{V}: Y_i(v)=0} f_I(s(X_i), v)$  and  
 123      $\gamma_i = \max_{v \in \mathcal{V}: Y_i(v)=1} f_O(-s(X_i), v)$  be the maxima of the function transformed scores over the  
 124     areas at which the true labels equal 0 and 1 respectively. We will require the following assumption  
 125     on the scores and the transformation functions.  
 126

127     **Assumption 2.** (Independence of scores)  $(X_i, Y_i)_{i=1}^{n+1}$  is independent of the functions  $s, f_O, f_I$ .  
 128

Given this we construct confidence sets as follows.

**Theorem 2.1.** (*Marginal inner set*) Under Assumptions 1 and 2, given  $\alpha_1 \in (0, 1)$ , let

$$131 \quad \lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1[\tau_i \leq \lambda] \geq \frac{\lceil (1 - \alpha_1)(n + 1) \rceil}{n} \right\},$$

134     and define  $I(X) = \{v \in \mathcal{V} : f_I(s(X), v) > \lambda_I(\alpha_2)\}$ . Then,

$$136 \quad \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1. \quad (3)$$

137     *Proof.* Under Assumptions 1 and 2, exchangeability of the image pairs implies exchangeability  
 138     of the sequence  $(\tau_i)_{i=1}^{n+1}$ . In particular, as  $\lambda_I(\alpha_1)$  is the upper  $\alpha_1$  quantile of the distribution of  
 139      $(\tau_i)_{i=1}^n \cup \{\infty\}$  by Lemma 1 of Tibshirani et al. (2019), it follows that  
 140

$$141 \quad \mathbb{P}(\tau_{n+1} \leq \lambda_I(\alpha_1)) \geq 1 - \alpha_1.$$

142     Now consider the event that  $\tau_{n+1} \leq \lambda_I(\alpha)$ . On this event,  $f_I(s(X_{n+1}), v) \leq \lambda_I(\alpha)$  for all  $v \in \mathcal{V}$   
 143     such that  $Y_{n+1}(v) = 0$ . As such, given  $u \in \mathcal{V}$  such that  $f_I(s(X_{n+1}), u) > \lambda_I(\alpha)$ , we must have  
 144      $Y_{n+1}(u) = 1$  so it follows that  $I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}$  and in particular that  
 145

$$146 \quad \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq \mathbb{P}(\tau_{n+1} \leq \lambda_I(\alpha_1)) \geq 1 - \alpha_1.$$

147                          □

149     For the outer set we have the following analogous result.

150     **Theorem 2.2.** (*Marginal outer set*) Under Assumptions 1 and 2, given  $\alpha_2 \in (0, 1)$ , let

$$152 \quad \lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1[\gamma_i \leq \lambda] \geq \frac{\lceil (1 - \alpha_2)(n + 1) \rceil}{n} \right\},$$

155     and define  $O(X) = \{v \in \mathcal{V} : f_O(-s(X), v) \leq \lambda_O(\alpha_2)\}$ . Then,

$$157 \quad \mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha_2. \quad (4)$$

159     *Proof.* Arguing as in the proof of Theorem 2.1, it follows that  $\mathbb{P}(\gamma_{n+1} \leq \lambda_O(\alpha_2)) \geq 1 - \alpha_2$ .  
 160     Now on the event that  $\gamma_{n+1} \leq \lambda_O(\alpha_2)$  we have  $f_O(-s(X_{n+1}, v)) \leq \lambda_O(\alpha_2)$  for all  $v \in \mathcal{V}$  such  
 161     that  $Y_{n+1}(v) = 1$ . As such, given  $u \in \mathcal{V}$  such that  $f_O(-s(X_{n+1}, u)) > \lambda_O(\alpha_2)$ , we must have  
 162      $Y_{n+1}(u) = 0$  and so  $O(X)^C \subseteq \{v \in \mathcal{V} : Y(v) = 0\}$ . The result then follows as above.                  □

**Remark 2.3.** We have used the maximum over the transformed scores in order to combine score information on and off the ground truth masks. The maximum is a natural combination function in imaging and is commonly used in the context of multiple testing (Worsley et al., 1992; Bowring et al., 2019). However the theory above is valid for any increasing combination function. We show this in Appendix A.1 where we establish generalized versions of these results.

**Remark 2.4.** Inner and outer coverage can also be viewed as a special case of conformal risk control with an appropriate choice of loss function. We can thus instead establish coverage results as a corollary to risk control, see Appendix A.2 for details. This amounts to an alternative proof of the results as the proof of the validity of risk control is different though still strongly relies on exchangeability.

### 2.3 JOINT CONFIDENCE SETS

Instead of focusing on marginal control one can instead spend all of the  $\alpha$  available to construct sets which have a joint probabilistic guarantees. This gain comes at the expense of a loss of precision. The simplest means of constructing jointly valid confidence sets is via the marginal sets themselves.

**Corollary 2.5.** (Joint from marginal) Assume Assumptions 1 and 2 hold and given  $\alpha \in (0, 1)$  and  $\alpha_1, \alpha_2 \in (0, 1)$  such that  $\alpha_1 + \alpha_2 \leq \alpha$ , define  $I(X)$  and  $O(X)$  as in Theorems 2.1 and 2.2. Then

$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq \frac{\lceil(1 - \alpha)(n + 1)\rceil}{n}. \quad (5)$$

Alternatively joint control can be obtained using the joint distribution of the maxima of the logit scores as follows.

**Theorem 2.6.** (Joint coverage) Assume that Assumption 1 and 2 hold. Given  $\alpha \in (0, 1)$ , define

$$\lambda(\alpha) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\max(\tau_i, \gamma_i) \leq \lambda] \geq 1 - \alpha \right\}.$$

Let  $O(X) = \{v \in \mathcal{V} : f_O(-s(X), v) \leq \lambda(\alpha)\}$  and  $I(X) = \{v \in \mathcal{V} : f_I(s(X), v) > \lambda(\alpha)\}$ . Then,

$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha. \quad (6)$$

*Proof.* Exchangeability of the image pairs implies exchangeability of the sequence  $(\tau_i, \gamma_i)_{i=1}^{n+1}$ . Moreover on the event that  $\max(\tau_{n+1}, \gamma_{n+1}) \leq \lambda(\alpha)$  we have  $\tau_{n+1} \leq \lambda(\alpha)$  and  $\gamma_{n+1} \leq \lambda(\alpha)$  so the result follows via a proof similar to that of Theorem 2.1.  $\square$

**Remark 2.7.** The advantage of Corollary 2.5 is that the resulting inner and outer sets provide pivotal inference - not favouring one side or the other - which can be important when the distribution of the score function is asymmetric. Moreover the levels  $\alpha_1$  and  $\alpha_2$  can be used to provide a greater weight to either inner or outer sets whilst maintaining joint coverage. Theorem 2.6 may instead be useful when there are strong levels of dependence between  $\tau_{n+1}$  and  $\gamma_{n+1}$ . However, when this dependence is low, scale differences in the scores can lead to a lack of pivotality. This can be improved by appropriate choices of the score transformations  $f_I$  and  $f_O$  however in practice it may be simpler to construct joint sets using Corollary 2.5.

### 2.4 BETTER SEGMENTORS PROVIDE MORE PRECISE CONFORMAL CONFIDENCE SETS

Given two real random variables,  $A$  and  $B$  write  $A \succeq B$  to indicate that  $\mathbb{P}(A > t) \geq \mathbb{P}(B > t)$  for all  $t \in \mathbb{R}$ . Then we have the following result.

**Theorem 2.8.** Suppose that  $(X_i, Y_i)_{i=1}^{n+1}$  is an i.i.d. sequence, and let  $s, t : \mathcal{V} \rightarrow \mathbb{R}$  be two score functions. Assume that  $\max_{v \in \mathcal{V}: Y_1(v)=0} s_v(X_1) \succeq \max_{v \in \mathcal{V}: Y_1(v)=0} t_v(X_1)$

### 2.5 OPTIMIZING SCORE TRANSFORMATIONS

The choice of score transformations  $f_I$  and  $f_O$  is extremely important and can have a large impact on the size of the conformal confidence sets. The best choice depends on both the distribution of the data and on the nature of the output of the trained segmentor used to calculate the scores. We thus

216 recommend setting aside a learning dataset independent from both the calibration dataset, used to  
 217 compute the conformal thresholds, and the test dataset. This approach was used in Sun & Yu (2024)  
 218 to learn the best copula transformation for combining dependent data streams.  
 219

220 In order to make efficient use of the data available, the learning dataset can in fact contain some  
 221 or all of the data used to train the image segmentor. This data is assumed to be independent of  
 222 the calibration and test data and so can be used to learn the best score transformations without  
 223 compromising validity . The advantage of doing so is that less additional data needs to be set aside  
 224 or collected for the purposes of learning a score function. Moreover it allows for additional data to  
 225 be used to train the model resulting in better segmentation performance. The disadvantage is that  
 226 machine learning models typically overfit their training data meaning that certain score functions  
 227 may appear to perform better on this data than they do in practice. The choice of whether to include  
 228 training data in the learning dataset thus depends on the quantity of data available and the quality of  
 229 the segmentation model.

230 A score transformation that we will make particular use of in Section 3 is based on the distance  
 231 transformation which we define as follows. Given  $\mathcal{A} \subseteq \mathcal{V}$ , let  $E(\mathcal{A})$  be the set of points on the  
 232 boundary of  $\mathcal{A}$  obtained using the marching squares algorithm (Maple, 2003). Given a distance  
 233 metric  $\rho$  define the distance transformation  $d_\rho : \mathcal{P}(\mathcal{V}) \times \mathcal{V} \rightarrow \mathbb{R}$ , which sends  $\mathcal{A} \in \mathcal{P}(\mathcal{V})$  and  $v \in \mathcal{V}$   
 234 to

$$d_\rho(\mathcal{A}, v) = \text{sign}(\mathcal{A}, v) \min\{\rho(v, e) : e \in E(\mathcal{A})\},$$

235 where  $\text{sign}(\mathcal{A}, v) = 1$  if  $v \in \mathcal{A}$  and equals  $-1$  otherwise. The function  $d_\rho$  is an adaption of the  
 236 distance transform of Borgefors (1986) which provides positive values within the set  $\mathcal{A}$  and negative  
 237 values outside of  $\mathcal{A}$ .  
 238

## 239 2.6 CONSTRUCTING CONFIDENCE SETS FROM BOUNDING BOXES

240 Existing work on conformal confidence sets which aim to provide coverage of the entire ground  
 241 truth mask with a given probability has primarily focused on bounding boxes, see e.g. (de Grancey  
 242 et al., 2022; Andéol et al., 2023; Mukama et al., 2024). These papers adjust for multiple comparisons  
 243 over the 4 edges of the bounding box, doing so conformally by comparing the distance between the  
 244 predicted bounding box and the bounding box of the ground truth mask. These approaches aggregate  
 245 the predictions over all objects within all of the calibration images, often combining multiple  
 246 bounding boxes per image. However, as observed in Section 5 of de Grancey et al. (2022), doing so  
 247 violates exchangeability which is needed for valid conformal inference, as there is dependence be-  
 248 tween the objects within each image. These papers do not provide formal proofs and their theoretical  
 249 validity is thus unclear.

250 In order to provide a more formal justification of bounding box methods we establish the validity  
 251 of an adapted version of the max-additive method of Andéol et al. (2023) as a corollary to our  
 252 results, see Appendix A.3. We compare to this approach in our experiments below. Targetting  
 253 bounding boxes does not directly target the mask itself and so the resulting confidence sets are  
 254 typically conservative.  
 255

## 256 3 APPLICATION TO POLPPS TUMOR SEGMENTATION

257 In order to illustrate and validate our approach we consider the problem of polyps tumor segmen-  
 258 tation. To do so we use the same dataset as in Angelopoulos et al. (2022) in which 1798 polyps  
 259 images, with available ground truth masks were combined from 5 open-source datasets (published  
 260 in Pogorelov et al. (2017), Borgli et al. (2020) Bernal et al. (2012), Silva et al. (2014)). Logit scores  
 261 were obtained for this data using the PraNet model Fan et al. (2020), which is based on the Unet  
 262 architecture CITE CHECK!

### 263 3.1 CHOOSING A SCORE TRANSFORMATION

264 In order to optimize the size of our confidence sets we set aside 298 of the 1798 polyps images  
 265 to form a learning dataset on which to choose the best score transformations. Importantly as the  
 266 learning dataset is independent of the remaining 1500 images set-aside, we can study it as much as  
 267 we like without compromising the validity of the follow-up analyses in Sections 3.2. In particular

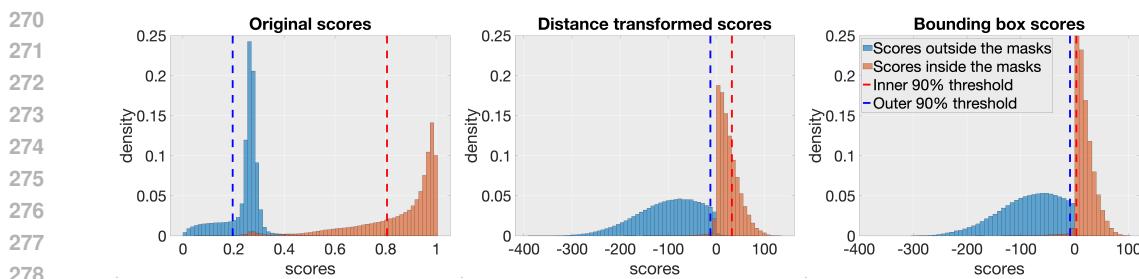


Figure 1: Histograms of the distribution of the scores on the learning dataset within and outside the ground truth masks. Thresholds obtained for the marginal 90% inner and outer confidence sets, based on the respective scores, are displayed in red and blue.

in this section we shall use the learning dataset as both calibrate and study the results, in order to maximize the amount of important information we can learn from it.

The score transformations we considered were the identity (after softmax transformation) and distance transformations of the predicted masks: taking  $f_I(s(X), v) = f_O(s(X), v) = d_\rho(\hat{M}(X), v)$ , where  $\rho$  is the Euclidean metric. We also compare to the results of using the bounding box transformations  $f_I = b_I$  and  $f_O = b_O$  which correspond to transforming the predicted bounding box using a distance transformation based on the chessboard metric and are defined formally in Appendix A.3. For the purposes of plotting we used the combined bounding box scores defined in Definition A.4.

From the histograms in Figure 1 we can see that thresholding the original scores at the inner threshold captures most of the data. However this is not the case for the outer threshold for which the data is better separated using the distance transformed and bounding box scores. Figure 2 shows PraNet scores for 2 typical examples, along with surface plots of the transformed scores and corresponding marginal confidence regions (with thresholds obtained from calibrating over the learning dataset). From these we see that PraNet often assigns a high softmax score to the polyps regions which decreases in the regions directly around the boundary of the tumor before returning to a higher level away from the polyps. This results in tight inner sets but large outer sets as the model struggles to identify where the tumor ends. Instead the distance transformed and bounding box scores are much better at providing outer bounds on the tumor, with distance transformed scores providing a tighter outside fit. Additional examples are shown in Figure 7 and 8 and have the same conclusion.

Based on the images set aside we can also learn the right balance of  $\alpha$  to use for joint confidence sets. We decided to use  $\alpha_1 = 0.02$  and  $\alpha_2 = 0.08$  to ensure a joint coverage of 90%. This ratio was chosen in light of the fact that in this dataset identifying where a given tumor ends appears to be more challenging than identifying pixels where we are sure that there is a tumor.

### 3.2 ILLUSTRATING THE PERFORMANCE OF CONFORMAL CONFIDENCE SETS

Based on the results of the learning dataset we decided to combine the best of the approaches for the inner and outer sets respectively, taking  $f_I$  to be the identity and  $f_O$  to be the distance transformation of the predicted mask.

We divide the 1500 images at random into 500 for conformal calibration, and 1000 for validation. The resulting conformal confidence sets for this data are shown in the second row of Figure 3. For comparison we have also shown the sets obtained based on the untransformed softmax scores in the top row. From this figure we see that the method, using the transformed scores, effectively delineates polyp regions. Inner sets are plotted in red and the outer sets are shown in blue. The ground truth mask for each polyps is shown in yellow and can be compared to the original images. In each of the examples considered the ground truth mask is bounded from within by the inner set and from without by the outer set.

The inner sets are shown in red and represent regions where we can have high confidence of the presence of polyps. The outer sets are shown in blue and represent regions in which the polyps may be.

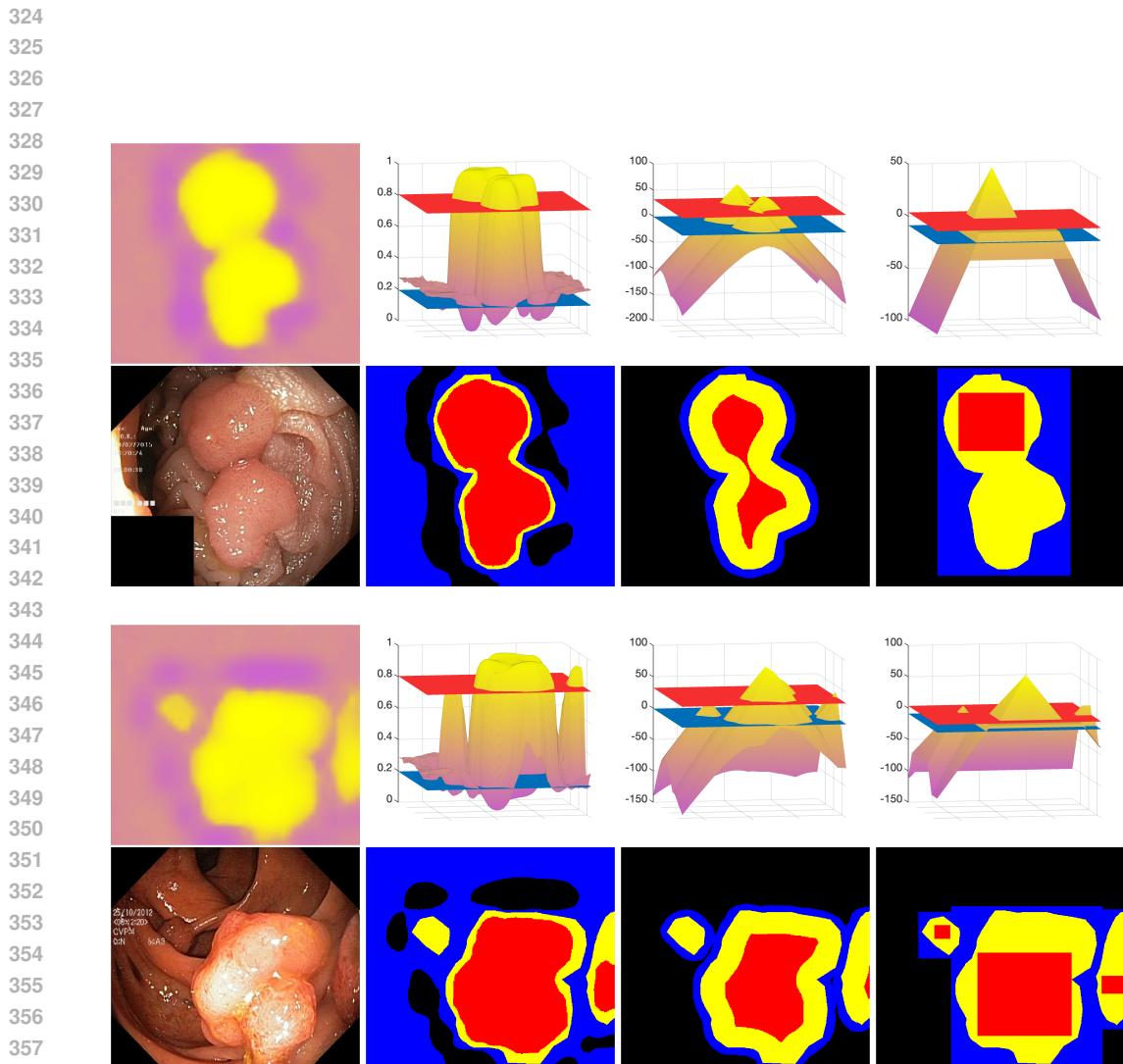


Figure 2: Illustrating the performance of the different score transformations on the learning dataset. We display 2 example tumors and present the results of each in 8 panels. These panels are as follows. Bottom right: the original image of the polyps tumor. Top Left: an intensity plot of the scores obtained from PraNet with purple/yellow indicating areas of lower/higher assigned probability. For the remaining panels, 3 different score transformations are shown which from left to right are the original scores, distance transformed scores  $d_\rho(\hat{M}(X), v)$  and bounding box scores (obtained using the combined bounding box score  $b_M$  defined in Definition A.4). In each of the panels on the top row a surface plot of the transformed PraNet scores is shown, along with the marginal conformal thresholds which are used to obtain the marginal 90% inner and outer sets. These thresholds are illustrated via red and blue planes respectively and are obtained over the learning dataset. The panels on the bottom show the corresponding conformal confidence sets. Here the inner set is shown in red, plotted over the ground truth mask of the polyps, shown in yellow, plotted over the outer set which is shown in blue. The outer set contains the ground truth mask which contains the inner set in all examples. From these figures we see that the original scores provide tight inner confidence sets and the distance transformed scores instead provide tight outer confidence sets. The conclusion from the learning dataset is therefore that it makes sense to combine these two score transformations.

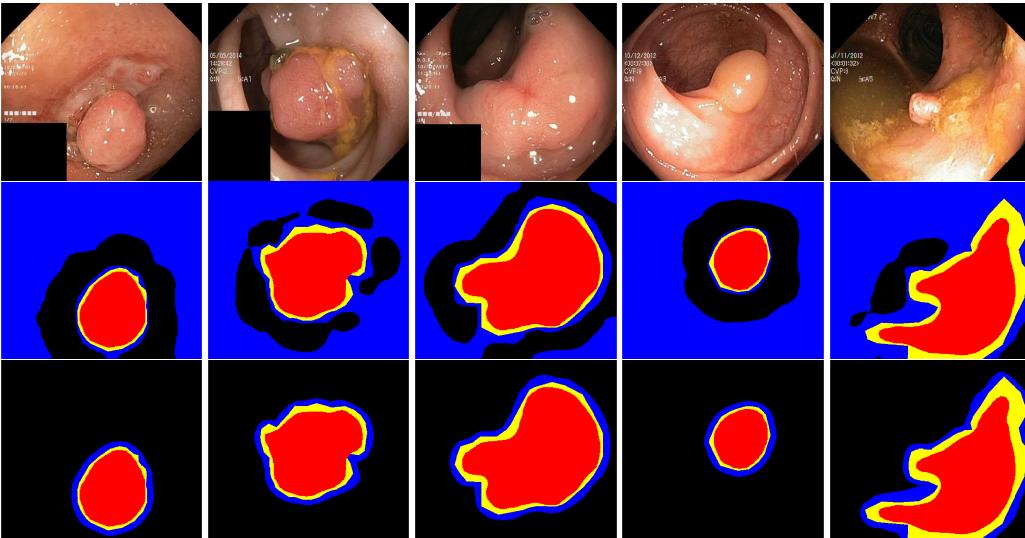


Figure 3: Conformal confidence sets for the polyps data. The bottom row shows the original endoscopic images with visible polyps. The top two rows present the conformal confidence sets, with the ground truth masks shown in yellow. The inner sets and outer sets are shown in red and blue respectively. The top row illustrates the sets which arise when using the original scores. Instead the middle show the resulting sets when  $f_O$  is given by the distance transformation of the predicted polyps mask. The figure shows the benefit of transforming the score function and illustrates the method’s effectiveness in accurately identifying polyp regions whilst providing informative spatial uncertainty bounds.

These results show that we can provide informative confidence bounds for the location of the polyps and allow us to use the PraNet segmentation model with uncertainty guarantees. They also illustrate the limitations of the model which is essential for applications. Larger uncertainty bounds may require specialist follow-up in order to be certain about the true extent of the observed tumor. Improved uncertainty quantification would require an improved segmentation model.

More precise results can be obtained at the expense of probabilistic guarantees, see Figure XXX. A trade off must be made between precision and confidence and this can also be determined in advance based on the learning dataset. The approach of CITE controls the empirical false negative risk yielding additional precision but at the cost of coverage as shown in Figure XXX.

### 3.3 MEASURING THE COVERGE RATE

In this section we run validations to evaluate the false coverage rate of our approach. To do so we take the set aside 1500 images and run 1000 validations, in each validation dividing the data into 1000 calibration and 500 test images. In each division we calculate the conformal confidence sets using the above approaches, based on thresholds derived from the calibration dataset, and evaluate the coverage rate on the test dataset. We average over all 1000 validations and present the results in Figure 4. Histograms for the 90% coverage obtained over each validation run are shown in Figure 9. From these results we can see that for all the approaches the coverage rate is controlled at or above the nominal level as desired. The coverage for the bounding box scores slightly over cover at lower levels. This is likely due to the discontinuities in the score functions.

XXX In this Figure we also compare to the coverage attained by using Conformal Risk control . We can see that conformal risk control can have highly inflated error rates - this is because it is designed to control the expected proportion of discoveries not cover the tumors. The results indicate the trade-off that must be made when choosing between the methodss, i.e. whilst risk control can provide meaningful inference CITE it comes with a cost in terms of under coverage. Instead, in this setting, conformal confidence sets provide informative segmentation bounds (as illustrated in Section 3.2) and come with strong coverage guarantees.

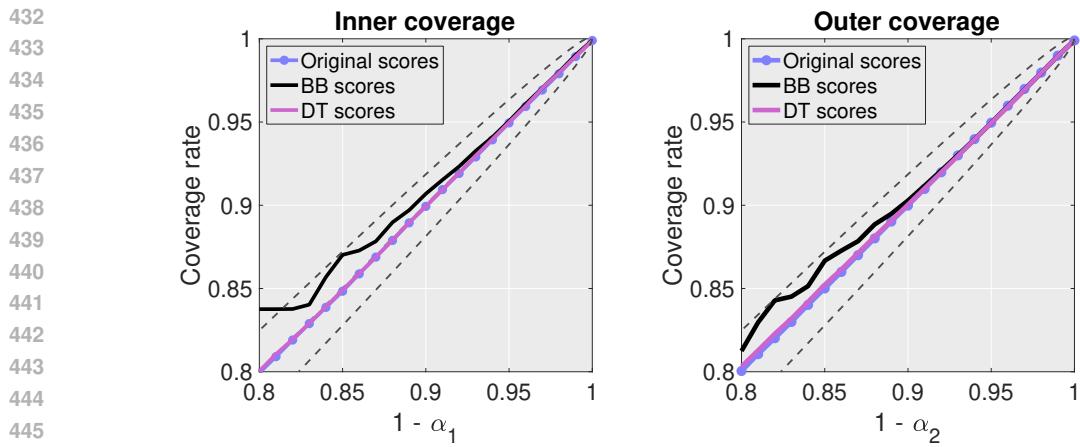


Figure 4: False coverage levels of the inner and outer sets averaged over 1000 validations for the original, distance transformed (DT) and bounding box (BB) scores.

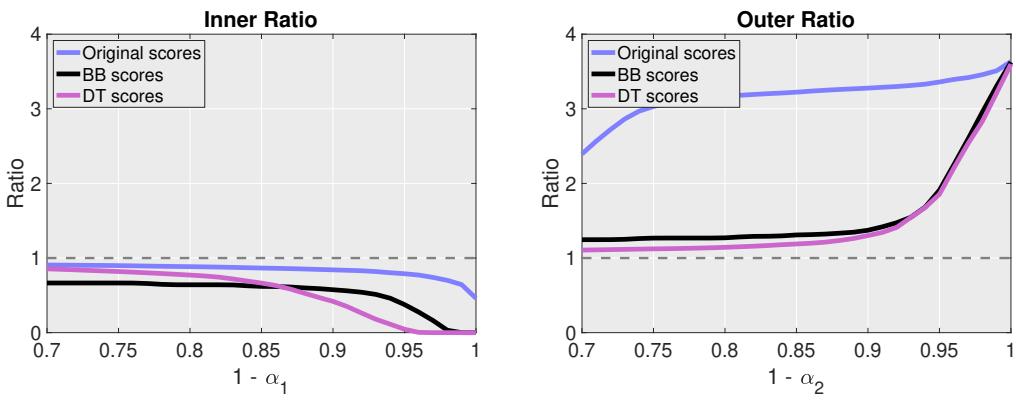


Figure 5: Measuring the efficiency of the bound using the ratio of the diameter of the coverage set to the diameter of the true tumor mask. The closer the ratio is to one the better. Higher coverage rates lead to a lower efficiency. The original scores provide the most efficient inner sets and the distance transformed scores provide the most efficient outer sets.

### 3.4 COMPARING THE EFFICIENCY OF THE BOUNDS

In order to compare the power of the approaches we compare the ratio of the

It follows that the method chosen based on the learning dataset which uses the distance transformed scores for the outer set and the original scores for the inner set is the best combination, providing the most precise confidence sets. Which matches the observations from Section 3.2.

### 3.5 IMPROVING RISK CONTROL USING TRANSFORMED SCORES

Risk control can also benefit

## 4 DISCUSSION

In this work, we have developed conformal confidence sets which offer probabilistic guarantees for the output of a image segmentation model. Our work helps to address the lack of formal uncertainty quantification in the application of deep neural networks to medical imaging which has limited the reliability and adoption of these models in practice.

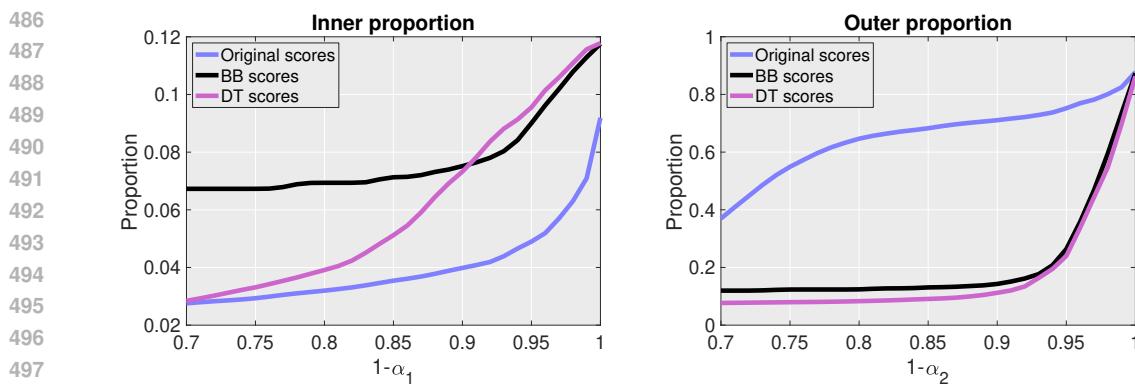


Figure 6: Measuring the proportion of the entire image which is under/over covered by the respective confidence sets. Left: proportion of the image which lies within the true mask but outside of the inner set. Middle: proportion of the image which lies within the confidence set but outside of the true mask. For both a lower proportion corresponds to increased precision.

Discuss how the method is very fast

One of the key strengths of our method is its ability to provide spatially resolved uncertainty estimates. Unlike global uncertainty measures, our approach allows for the identification of specific regions within an image where the model’s predictions are less certain.

Future work could explore more efficient algorithms or approximations that maintain the statistical guarantees while reducing computational cost. Second, while our method provides valid coverage guarantees, the tightness of the confidence sets may vary depending on the underlying model’s performance and the complexity of the segmentation task. In some cases, the confidence sets may be conservatively large, potentially limiting their practical utility. Investigating ways to produce tighter confidence sets while maintaining coverage guarantees is an important direction for future research.

Third, our current approach treats each pixel or voxel independently when constructing confidence sets. This may not fully capture the spatial correlations inherent in many biological structures. Developing methods that incorporate spatial dependencies and prior anatomical knowledge could lead to more informative and biologically plausible uncertainty estimates.

The implications of our work extend beyond the immediate technical contributions. By providing a rigorous framework for uncertainty quantification, we address a critical need in the deployment of AI systems in high-stakes applications like medical diagnosis. Our method can enhance the trustworthiness of AI-assisted image analysis by clearly communicating the limits of model certainty. This transparency is crucial for responsible AI deployment and could help mitigate risks associated with overreliance on automated systems.

Moreover, the insights gained from our uncertainty estimates could feed back into the development of improved segmentation models. By identifying consistent patterns of uncertainty, researchers may uncover systematic limitations in current architectures or training approaches, guiding future innovations in the field.

In conclusion, our work represents a significant step forward in bringing the power of conformal prediction to the domain of image segmentation. By providing spatial uncertainty guarantees with finite sample validity, we offer a valuable tool for researchers and clinicians alike. As AI continues to play an increasingly prominent role in medical imaging and beyond, methods like ours will be essential in ensuring that these powerful technologies are deployed responsibly and effectively.

Additionally, investigating the relationship between model calibration, uncertainty estimates, and out-of-distribution detection could further enhance the robustness of AI systems in real-world deployment scenarios.

540 Our approach has the potential to help enhance the overall reliability and trustworthiness of AI-  
 541 assisted image analysis systems. By clearly delineating the limits of model certainty, we can help  
 542 prevent overconfidence in automated predictions and promote a more nuanced integration of AI  
 543 tools into professional workflows.

## 545 AVAILABILITY OF CODE

546 Matlab code to reproduce the results of the paper is available in the supplementary material.

## 549 REFERENCES

- 551 Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection  
 552 via conformal prediction and conformal risk control: an application to railway signaling. In  
 553 *Conformal and Probabilistic Prediction with Applications*, pp. 36–55. PMLR, 2023.
- 555 Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and  
 556 distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.
- 557 Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua  
 558 Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint*  
 559 *arXiv:2110.01052*, 2021.
- 561 Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal  
 562 risk control. *arXiv preprint arXiv:2208.02814*, 2022.
- 563 Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan.  
 564 Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.
- 566 Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful  
 567 approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*,  
 568 57(1):289–300, 1995.
- 569 Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a  
 570 polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.
- 572 Gunilla Borgefors. Distance transformations in digital images. *Computer vision, graphics, and*  
 573 *image processing*, 34(3):344–371, 1986.
- 574 Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland,  
 575 Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al.  
 577 HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy.  
*Scientific data*, 7(1):283, 2020.
- 579 Alexander Bowring, Fabian Telschow, Armin Schwartzman, and Thomas E. Nichols. Spatial confi-  
 580 dence sets for raw effect size images. *NeuroImage*, 203:116187, 2019.
- 581 Florence de Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and  
 582 David Vigouroux. Object detection with probabilistic guarantees. In *Fifth International Workshop*  
 583 *on Artificial Intelligence Safety Engineering (WAISE 2022)*, 2022.
- 585 Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao.  
 586 Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on*  
 587 *medical image computing and computer-assisted intervention*, pp. 263–273. Springer, 2020.
- 588 Weikang Gong, Lin Wan, Wenlian Lu, Liang Ma, Fan Cheng, Wei Cheng, Stefan Gruenewald, and  
 589 Jianfeng Feng. Statistical testing and power analysis for brain-wide association study. *Medical*  
 590 *image analysis*, 47:15–30, 2018.
- 592 Carsten Maple. Geometric design and space planning using the marching squares and marching  
 593 cube algorithms. In *2003 international conference on geometric modeling and graphics, 2003.*  
*Proceedings*, pp. 90–95. IEEE, 2003.

594 Bruce Cyusa Mukama, Soundouss Messoudi, Sylvain Rousseau, and Sébastien Destercke. Copula-  
 595 based conformal prediction for object detection: a more efficient approach. *Proceedings of Ma-*  
 596 *chine Learning Research*, 230:1–18, 2024.

597 Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas  
 598 de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Pe-  
 599 ter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset  
 600 for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multi-*  
 601 *media Systems Conference*, MMSys’17, pp. 164–169, New York, NY, USA, 2017. ACM. ISBN  
 602 978-1-4503-5002-0. doi: 10.1145/3083187.3083212.

603 Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward em-  
 604 bedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International*  
 605 *journal of computer assisted radiology and surgery*, 9:283–293, 2014.

606 Sophia Sun and Rose Yu. Copula conformal prediction for multi-step time series forecasting. In  
 607 *International Conference on Learning Representations (ICLR)*, 2024.

608 Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal pre-  
 609 diction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

610 Keith J. Worsley, Alan C Evans, Sean Marrett, and P Neelin. A three-dimensional statistical analysis  
 611 for CBF activation studies in human brain. *JCBFM*, 1992.

## 612 A APPENDIX

### 613 A.1 OBTAINING CONFORMAL CONFIDENCE SETS WITH INCREASING COMBINATION 614 FUNCTIONS

615 As discussed in Remark 2.3 the results of Sections 2.2 and 2.3 can be generalized to a wider class  
 616 of combination functions.

617 **Definition A.1.** We define a suitable combination function to be a function  $C : \mathcal{P}(\mathcal{V}) \times \mathcal{X} \rightarrow \mathbb{R}$   
 618 which is increasing in the sense that for all sets  $\mathcal{A} \subseteq \mathcal{V}$  and each  $v \in \mathcal{A}$ ,  $C(v, X) \leq C(\mathcal{A}, X)$  for  
 619 all  $X \in \mathcal{X}$ .

620 The maximum is a suitable combination function since  $X(v) = \max_{v \in \{v\}} X(v) \leq \max_{v \in \mathcal{A}} X(v)$ .  
 621 As such this framework directly generalizes the results of the main text.

622 We can construct generalized marginal confidence sets as follows.

623 **Theorem A.2.** (*Marginal inner set*) Under Assumptions 1 and 2, given  $\alpha_1 \in (0, 1)$ , define

$$624 \lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n \mathbb{1}[C(\{v \in \mathcal{V} : Y_i(v) = 1\}, f_I(s(X_i))) \leq \lambda] \geq 1 - \alpha_1 \right\},$$

625 for a suitable combination function  $C$ , and define  $I(X) = \{v \in \mathcal{V} : C(v, f_I(s(X))) > \lambda_I(\alpha_1)\}$ .  
 626 Then,

$$627 \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1} = 1\}) \geq 1 - \alpha_1. \quad (7)$$

628 The proof follows that of Theorem 2.1. The key observation is that for any suitable combination  
 629 function  $C$ , given  $\lambda \in \mathbb{R}$ ,  $\mathcal{A} \subseteq \mathcal{V}$  and  $X \in \mathcal{X}$ , we have that  $C(\mathcal{A}, X) \leq \lambda$  implies that  $C(v, X) \leq \lambda$ .  
 630 This is the relevant property of the maximum which we used for the results in the main text. For the  
 631 outer set we similarly have the following.

632 **Theorem A.3.** (*Marginal outer set*) Under Assumptions 1 and 2, given  $\alpha_2 \in (0, 1)$ , define

$$633 \lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n \mathbb{1}[C(\{v \in \mathcal{V} : Y_i(v) = 0\}, f_O(-s(X_i))) \leq \lambda] \geq 1 - \alpha_2 \right\}.$$

634 for a suitable combination function  $C$ , and let  $O(X) = \{v \in \mathcal{V} : C(v, f_O(-s(X))) \leq \lambda_O(\alpha_2)\}$ .  
 635 Then,

$$636 \mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha_2. \quad (8)$$

637 Joint results can be analogously obtained.

648 A.2 OBTAINING CONFIDENCE SETS FROM RISK CONTROL  
 649

650 We can alternatively establish Theorems 2.1 and A.2 using an argument from risk control (An-  
 651 gelopoulos et al., 2022). In particular, given an image pair  $(X, Y)$  and  $\lambda \in \mathbb{R}$ , let

652 
$$I_\lambda(X) = \{v \in \mathcal{V} : C(v, f_I(s(X))) > \lambda\}.$$
  
 653

654 Define a loss function,  $L : \mathcal{P}(\mathcal{V}) \times \mathcal{Y} \rightarrow \mathbb{R}$  which sends  $(X, Y)$  to  
 655

656 
$$L(I_\lambda(X), Y) = 1 [I_\lambda(X) \not\subseteq \{v \in \mathcal{V} : Y_{n+1} = 1\}].$$

657 For  $i = 1, \dots, n+1$ , let  $L_i(\lambda) = L(I_\lambda(X_i), Y_i)$ . Then applying Theorem 1 of Angelopoulos et al.  
 658 (2022) it follows that

659 
$$\mathbb{E} [L_{n+1}(\hat{\lambda})] \leq \alpha_1$$
  
 660

661 where  $\hat{\lambda} = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n L_i(\lambda) \leq \alpha_1 - \frac{1-\alpha_1}{n} \right\}$ . Arguing as in Appendix A of (Angelopoulos  
 662 et al., 2022) it in fact follows that  $\hat{\lambda} = \lambda_I(\alpha_1)$  and so  $I(X) = I_{\hat{\lambda}}(X)$ . As such

663 
$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1} = 1\}) = 1 - \mathbb{E} [L_{n+1}(\hat{\lambda})] \geq 1 - \alpha_1, \quad (9)$$
  
 664

665 and we recover the desired result. Arguing similarly it is possible to establish proofs of Theorems  
 666 2.2 and A.3.

667 A.3 PROVIDING THEORY FOR DERIVING CONFIDENCE SETS FROM BOUNDING BOXES  
 668

669 We can use our results in order to provide valid inference for bounding boxes. In what follows we  
 670 adapt the approach of Andéol et al. (2023) in order to ensure validity. In particular given  $Z \in \mathcal{Y}$ , let  
 671  $B_{I,\max}(Z)$  be the largest box which can be contained within the set  $\{v \in \mathcal{V} : Z(v) = 1\}$  and let  
 672  $B_{O,\min}(Z)$  be the smallest box which contains it. Given  $Y \in \mathcal{Y}$ , let  $cc(Y) \subseteq \mathcal{P}(\mathcal{V})$  denote the set  
 673 of connected components of the set  $\{v \in \mathcal{V} : Y(v) = 1\}$  for a given connectivity criterion (which  
 674 we take to be 4 in our examples), and note that these can themselves be identified as elements of  $\mathcal{Y}$ .  
 675 Define

676 
$$B_I(Y) = \cup_{c \in cc(Y)} B_{I,\max}(c) \text{ and } B_O(Y) = \cup_{c \in cc(Y)} B_{O,\min}(c)$$
  
 677

678 to be the unions of the largest inner and smallest outer boxes of the connected components of the  
 679 image  $Y$ , respectively. Then define

680 
$$\hat{B}_I(s(X)) = \cup_{c \in cc(\hat{M}(X))} B_{I,\max}(c) \text{ and } \hat{B}_O(s(X)) = \cup_{c \in cc(\hat{M}(X))} B_{O,\min}(c)$$
  
 681

682 to be the unions of the largest inner and smallest outer boxes of the connected components of the  
 683 predicted mask  $\hat{M}(X)$ , respectively. Note that this is well-defined as  $\hat{M}(X)$  is a function of  $s(X)$ .

684 For the remainder of this section we shall assume that  $\mathcal{V} \subset \mathbb{R}^2$ , this is not strictly necessary but  
 685 will help to simplify notation. Given  $u, v \in \mathcal{V}$ , write  $u = (u_1, u_2)$  and  $v = (v_1, v_2)$  and let  
 686  $\rho(u, v) = \max(|u_1 - v_1|, |u_2 - v_2|)$  be the chessboard metric.

687 **Definition A.4.** (Bounding box scores) For each  $X \in \mathcal{X}$  and  $v \in \mathcal{V}$ , let

688 
$$b_I(s(X), v) = d_\rho(\hat{B}_I(s(X)), v) \text{ and } b_O(s(X), v) = d_\rho(\hat{B}_O(s(X)), v)$$
  
 689

690 be the distance transformed scores based on the chessboard distance to the predicted inner and outer  
 691 box collections  $\hat{B}_I(s(X))$  and  $\hat{B}_O(s(X))$ , respectively. We also define a combination of these  $b_M$ ,  
 692 primarily for the purposes of plotting in Figure 2, as follows. Let  $b_M(s(X), v) = b_O(s(X), v)$  for  
 693 each  $v \notin \hat{B}_O$  and let  $b_M(s(X), v) = \max(b_I(s(X), v), 0)$  for  $v \in \hat{B}_O$ . We shall write  $b_I(s(X)) \in$   
 694  $\mathcal{X}$  to denote the image which has  $b_I(s(X))(v) = b_I(s(X), v)$  and similarly for  $b_O(s(X))$  and  
 695  $b_M(s(X))$ . An illustration of these scores for two example tumors is shown in Figure XXX.

696 Now consider the sequences of image pairs  $(X_i, B_i^I)_{i=1}^n$  and  $(X_i, B_i^O)_{i=1}^n$ . These both satisfy ex-  
 697 changeability and so, applying Theorems 2.1 and 2.2 we obtain the following bounding box validity  
 698 results.

702     **Corollary A.5.** (*Marginal inner bounding boxes*) Suppose Assumption 1 holds and that  $(X_i, Y_i)_{i=1}^{n+1}$   
 703     is independent of the functions  $s$  and  $b_I$ . Given  $\alpha_1 \in (0, 1)$ , define  
 704

$$705 \quad \lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(B_i^I, b_I(s(X_i))) \leq \lambda] \geq \frac{\lceil (1 - \alpha_1)(n + 1) \rceil}{n} \right\}, \quad (10)$$

708     for a suitable combination function  $C$ , and define  $I(X) = \{v \in \mathcal{V} : C(v, b_I(s(X))) > \lambda_I(\alpha_1)\}$ .  
 709     Then,

$$710 \quad \mathbb{P}(I(X_{n+1}) \subseteq B_{n+1}^I \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1.$$

711     **Corollary A.6.** (*Marginal outer bounding boxes*) Suppose Assumption 1 holds and that  $(X_i, Y_i)_{i=1}^{n+1}$   
 712     is independent of the functions  $s$  and  $b_O$ . Given  $\alpha_2 \in (0, 1)$ , define  
 713

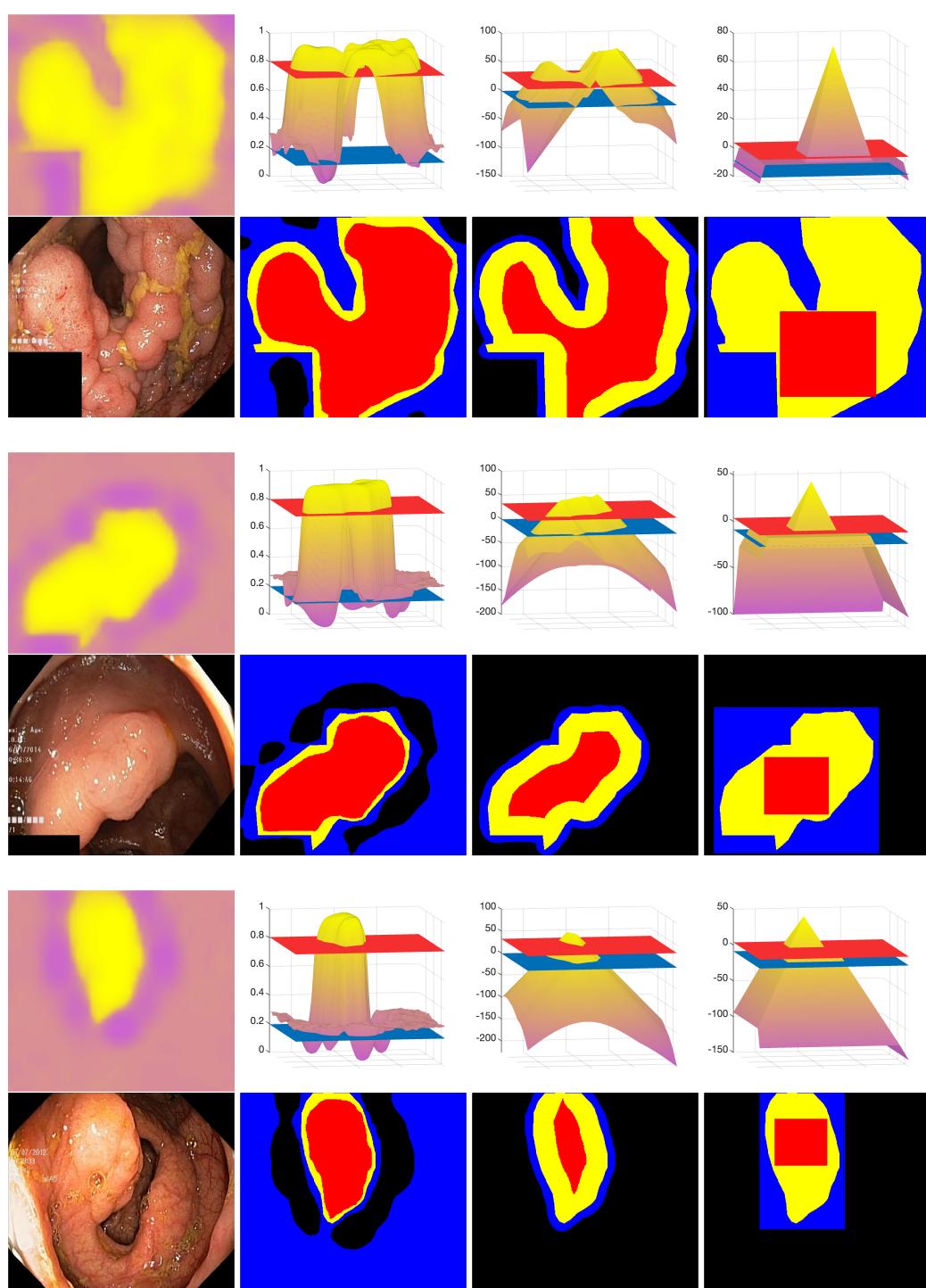
$$714 \quad \lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(B_i^O, -b_O(s(X_i))) \leq \lambda] \geq \frac{\lceil (1 - \alpha_2)(n + 1) \rceil}{n} \right\}. \quad (11)$$

717     for a suitable combination function  $C$ , and let  $O(X) = \{v \in \mathcal{V} : C(v, -b_O(s(X))) \leq \lambda_O(\alpha_2)\}$ .  
 718     Then,

$$719 \quad \mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq B_{n+1}^O \subseteq O(X_{n+1})) \geq 1 - \alpha_2.$$

720     Joint results can be obtained in a similar manner to those in Section 2.3.

721  
 722  
 723  
 724  
 725  
 726  
 727  
 728  
 729  
 730  
 731  
 732  
 733  
 734  
 735  
 736  
 737  
 738  
 739  
 740  
 741  
 742  
 743  
 744  
 745  
 746  
 747  
 748  
 749  
 750  
 751  
 752  
 753  
 754  
 755

756 A.4 ADDITIONAL EXAMPLES FROM THE LEARNING DATASET  
757804  
805 Figure 7: Additional examples from the learning dataset. The layout of these figures is the same as  
806 for Figure 2.  
807  
808  
809

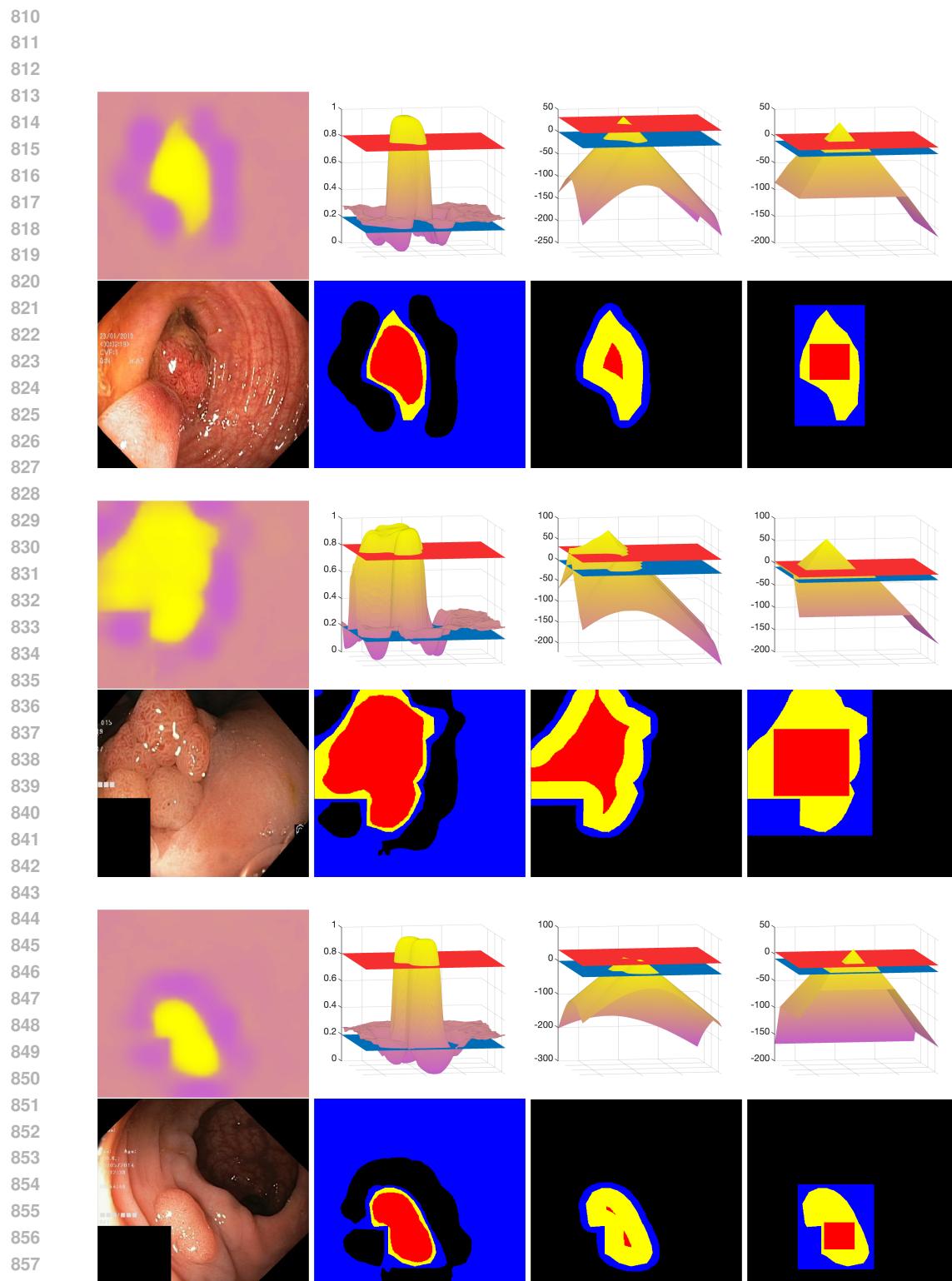
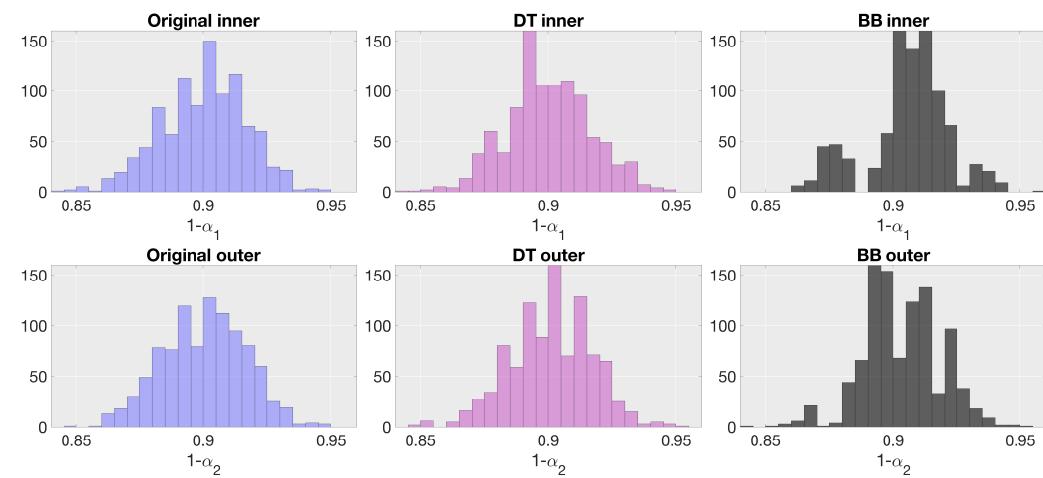


Figure 8: Futher examples from the learning dataset. The layout of these figures is the same as for Figure 2.

864 A.5 HISTOGRAMS OF THE COVERAGE  
865

882 Figure 9: Histograms of the coverage rates obtained across each of the validation resamples for  
883 90% inner and outer marginal confidence sets. We plot the results for the original scores, distance  
884 transformed scores (DT) and boundary box scores (BB) from left to right. The bounding box scores  
885 are discontinuous which is the cause of the discreteness of the rightmost histogram.  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

918

## A.6 ADDITIONAL EXAMPLES FROM THE VALIDATION SET

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971