# Supplemental Materials

## A Derivation of the SG-RCPS algorithm

*Without subgroups:* We start by summarizing the situation without subgroups. Let $X$ be the set of features and $Y$ the corresponding set of responses. Additionally, let $\mathcal{T}$ be a predictor with $\mathcal{T} : X \rightarrow \hat{Y}$ where $\hat{Y}$ is the space of sets that include different responses $Y$. The risk of $\mathcal{T}$ is defined as

$$\mathcal{R}(\mathcal{T}) = \mathbb{E}\left[\mathbb{1}_{\{Y \notin \mathcal{T}(X)\}}\right] = \Pr(Y \notin \mathcal{T}(X)),$$

where the expectation is taken over the distribution of $X$ and $Y$ on the calibration data set. Using the RCPS framework we can construct a predictor $\mathcal{T}$ such that $\mathcal{R}(\mathcal{T}) \leq \alpha$ with a probability of at least $1-\delta$. We note that the indicator function is bounded, which means that the risk is guaranteed to be bounded by

$$\begin{aligned}\mathcal{R}(\mathcal{T}) &= \mathbb{E}[\mathbb{1}_{\{Y \notin \mathcal{T}(X)\}}] \\ &= \mathbb{E}[\mathbb{1}_{\{Y \notin \mathcal{T}(X)\}} | \mathcal{R}(\mathcal{T}) \leq \alpha] \cdot \Pr(\mathcal{R}(\mathcal{T}) \leq \alpha) \\ &\quad + \mathbb{E}[\mathbb{1}_{\{Y \notin \mathcal{T}(X)\}} | \mathcal{R}(T) > \alpha] \cdot (1 - \Pr(\mathcal{R}(\mathcal{T}) \leq \alpha)) \\ &\leq \alpha + \delta.\end{aligned} \tag{1}$$

*With subgroups:* We model the case with multiple subgroups in the dataset by introducing an additional random variable $Z$ that takes values in $\{1, \ldots, K\}$ and addresses the different subgroups. For example, if $Z$ takes the value 1, $(X, Y)$ is assumed to be distributed according to the first subgroup, if $Z$ takes the value 2, $(X, Y)$ is distributed according to the second subgroup, etc. The value of $Z$ is unknown at test time. Algorithm 1 ensures that the risk for each subgroup is bounded via Eq. (1), that is,

$$\mathcal{R}(\mathcal{T}) = \Pr(Y \notin \mathcal{T}(X)) = \mathbb{E}[\mathbb{1}_{\{Y \notin \mathcal{T}(X)\}} | Z = \bar{z}] \leq \alpha + \delta,$$

for the distribution of $(X, Y)$ conditioned on the subgroup $\bar{z}$. This is due to the fact that Algorithm 1 applies the upper confidence bound arising from Hoeffding's inequality for each subgroup $\bar{z} \in \{1, \ldots K\}$ separately. The fact that the risk of the predictor $\mathcal{T}$ is bounded by $\alpha + \delta$ (conditional on $Z$), implies that the same holds for the distribution of $(X, Y)$ during test time:

$$\begin{aligned}\mathcal{R}(\mathcal{T}) = \Pr(Y \notin \mathcal{T}(X)) &= \mathbb{E}_{XY}\left[\mathbb{1}_{\{Y \notin \mathcal{T}(X)\}}\right] \\ &= \mathbb{E}_Z[\underbrace{\mathbb{E}_{XY|Z}\left[\mathbb{1}_{\{Y \notin \mathcal{T}(X)\}}\right]}_{\leq \alpha + \delta}] \\ &\leq \alpha + \delta.\end{aligned} \tag{2}$$

# B Summary of the dataset

**Table 1.** Number of patients and segments per tumour entity in training dataset

| | Training Dataset | | | |
|---|---|---|---|---|
| Overall Segments | 3958 | | | |
| Entity | Prostate | Liver | HN | Mamma |
| Patients | 40 | 15 | 15 | 5 |
| Segments | 2015 | 821 | 1013 | 109 |

**Table 2.** Number of patients and segments per tumour entity in test dataset

| | Test Dataset | | | | |
|---|---|---|---|---|---|
| Overall Segments | 2657 | | | | |
| Entity | Prostate | Liver | HN | Mamma | Lymphnodes |
| Patients | 10 | 10 | 10 | 5 | 15 |
| Segments | 646 | 525 | 731 | 128 | 627 |