Evolution of groups at high risk of death from Covid-19 using hospital data

Pierre-Yves Boëlle[1], Anna Bonnet[2], **Felix Cheysson**[2], Charlotte Dion[2], Olivier Lopez[2], Maud Thomas[2]

[1] Institut Pierre-Louis d'Epidémiologie et de Santé Publique
[2] Sorbonne Université, LPSM

EcoDep 2021 Conference
September 15th 2021

Outline

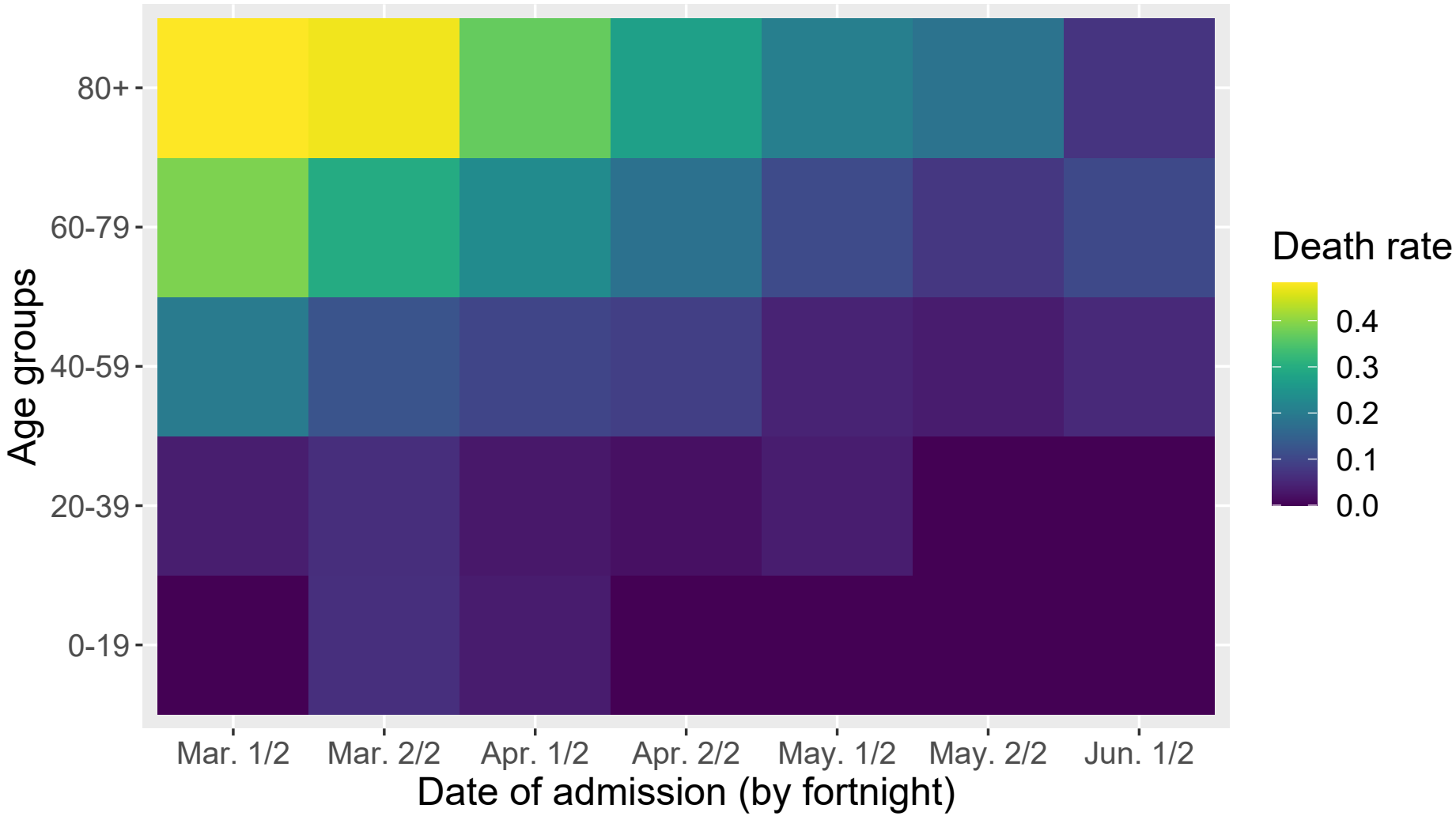
Some facts about the first wave

Overview of the pathway for hospitalised Covid-19 patients (Courtejoie and Dubost, 2020):

- SI-VIC database (*système d'information pour le suivi des victimes*) to monitor hospital admissions in the event of exceptional sanitary situations.
- Overall mortality rate: 19%; halved between early March and mid June.
- 17% for women, 21% for men; 2% for < 40 y.o., 33% for > 80 y.o.
- Median age for deceased individuals: 81 years.

Covid-19 death rates during first wave, Ile-de-France

Decrease in mortality across all age groups.



Modelling motivation

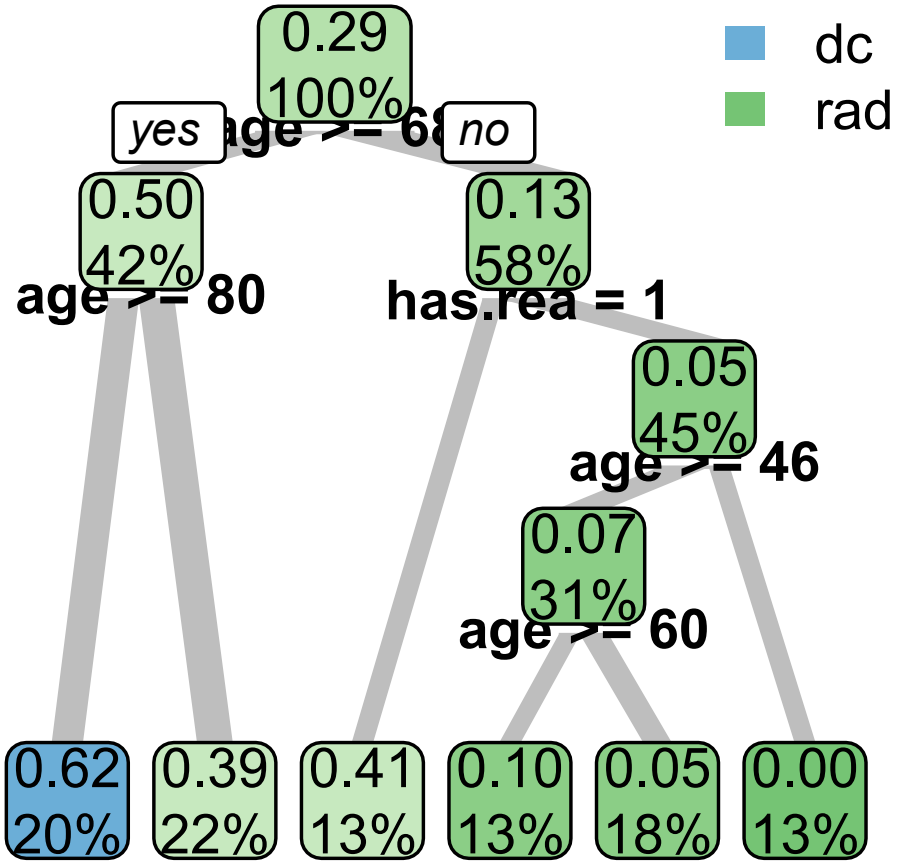
Covid-19 Dataset from SI-VIC database: all hospitalisation for Covid-19 patients in AP-HP hospitals.

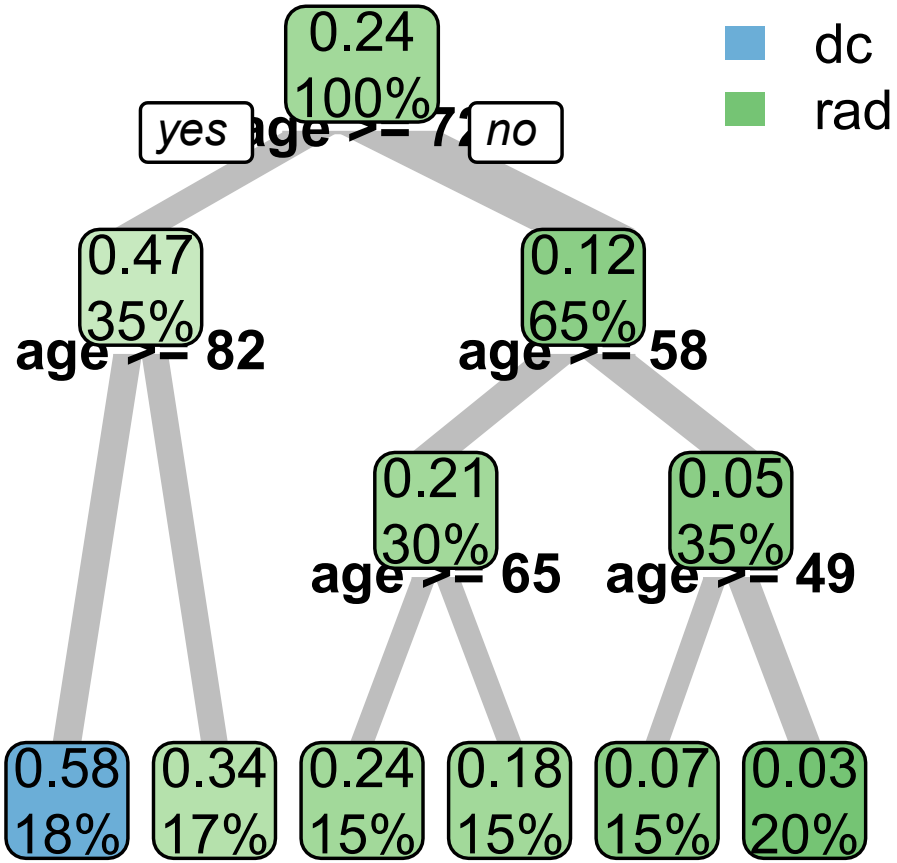
dt.first	dt.last	outcome	sex	age	hospital
2020-03-17	2020-04-05	rad	F	45	Robert Debré
2020-03-14	2020-03-25	rad	F	29	Robert Debré
2020-03-18	2020-03-29	dc	H	80	St Antoine
2020-03-11	2020-03-15	dc	H	62	St Louis
2020-03-04	2020-03-09	dc	F	72	Pitié Salpétrière
2020-03-16	2020-03-20	dc	H	92	Raymond Poincaré

- **Motivation**: We wish to model the risk of death of a patient hospitalised for Covid-19 with respect to covariates.
- **Objective**: Adapt care of patients when changes in the vulnerability of groups at risks are detected.

Estimating groups at risk using classification trees

- Classification and Regression Trees (Breiman et al., 1984)
- Build one classification tree per week.
- Study the evolution of mortality in groups at risk.





Classification and Regression Trees (CART)

A

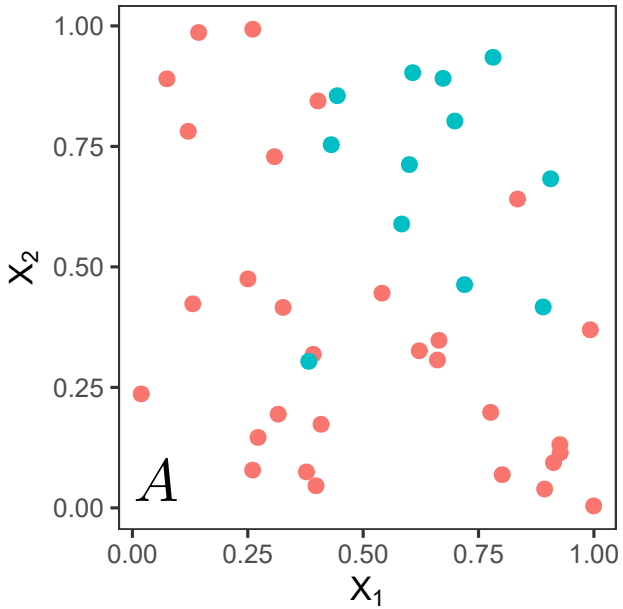
Introduced by Breiman et al., 1984.

- Construct binary tree by recursively splitting the sample space $\mathcal{X}$ along one of the covariate dimensions:
  - Find the dimension d and the value z such that the split (d, z) maximises the decrease in impurity:

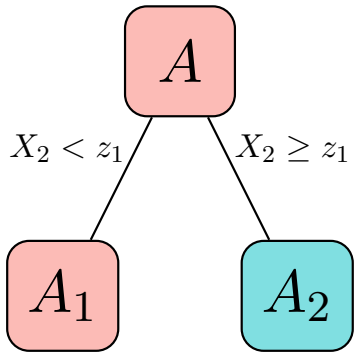
$$\Delta i(d, z) = i(A) - p_L i(A_L) - p_R i(A_R);$$

  - Label the node through majority vote;
  - Stop when a stopping rule is achieved.
- Prune the tree to reduce overfitting.

Extensions include randomised ensembles: random forests, bagging, etc.



Classification and Regression Trees (CART)



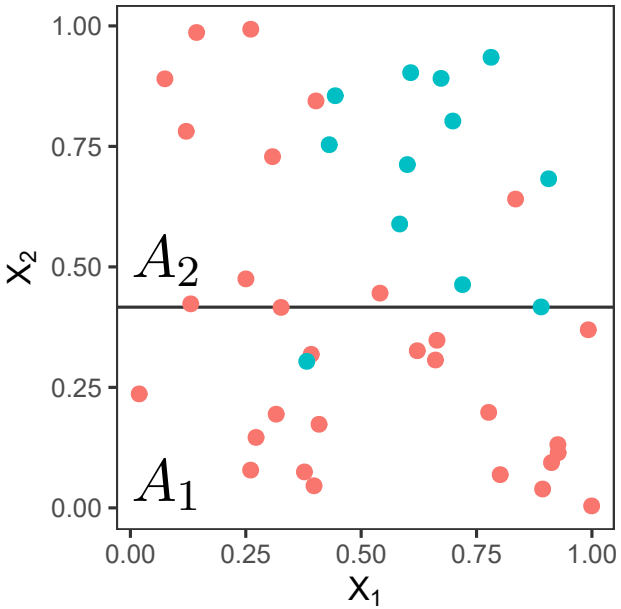
Introduced by Breiman et al., 1984.

- Construct binary tree by recursively splitting the sample space $\mathcal{X}$ along one of the covariate dimensions:
  - Find the dimension d and the value z such that the split (d, z) maximises the decrease in impurity:

$$\Delta i(d, z) = i(A) - p_L i(A_L) - p_R i(A_R);$$

  - Label the node through majority vote;
  - Stop when a stopping rule is achieved.
- Prune the tree to reduce overfitting.

Extensions include randomised ensembles: random forests, bagging, etc.



Classification and Regression Trees (CART)

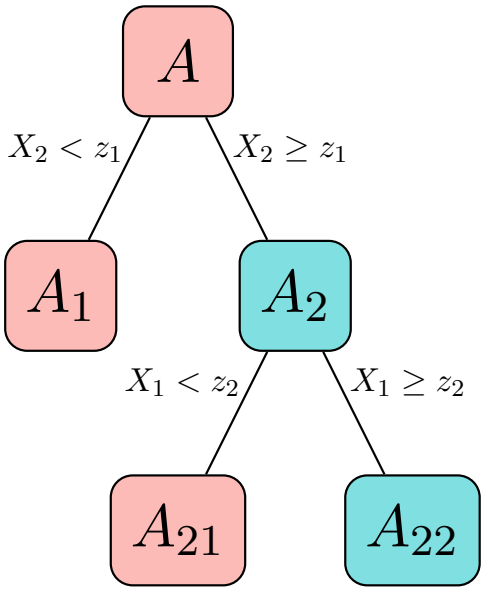
Introduced by Breiman et al., 1984.

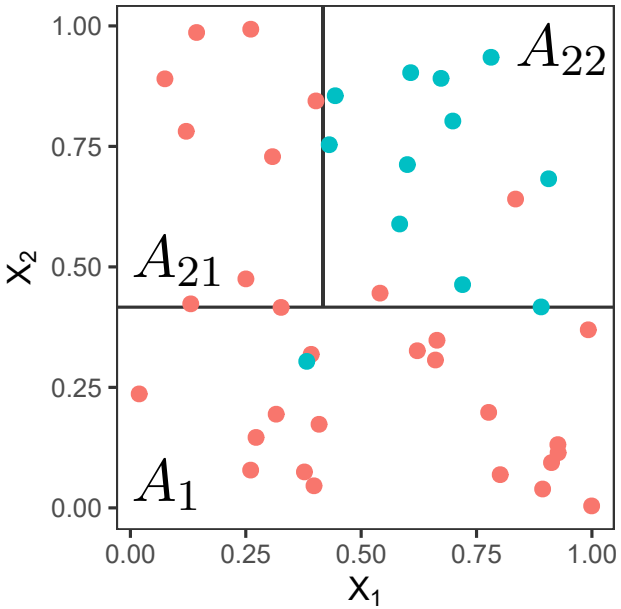
- Construct binary tree by recursively splitting the sample space $\mathcal{X}$ along one of the covariate dimensions:
  - Find the dimension d and the value z such that the split (d, z) maximises the decrease in impurity:

$$\Delta i(d, z) = i(A) - p_L i(A_L) - p_R i(A_R);$$

  - Label the node through majority vote;
  - Stop when a stopping rule is achieved.
- Prune the tree to reduce overfitting.

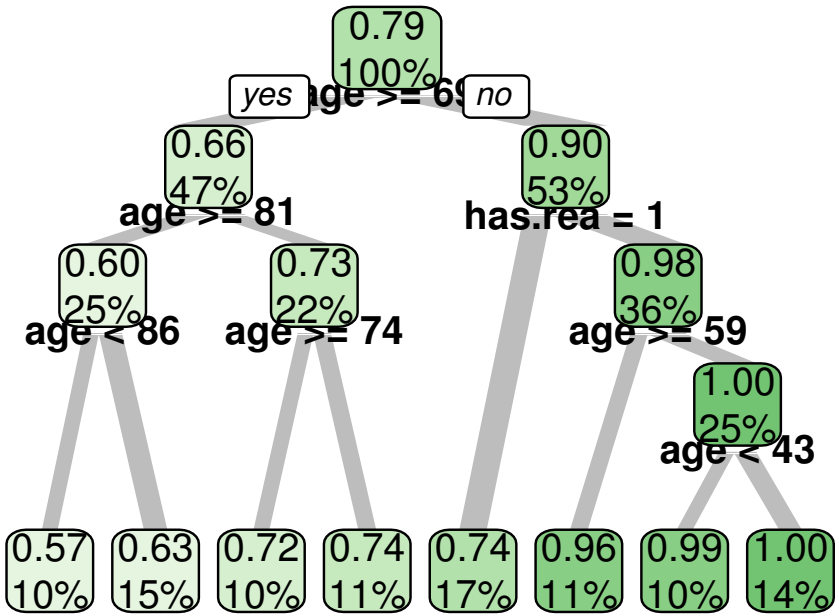
Extensions include randomised ensembles: random forests, bagging, etc.

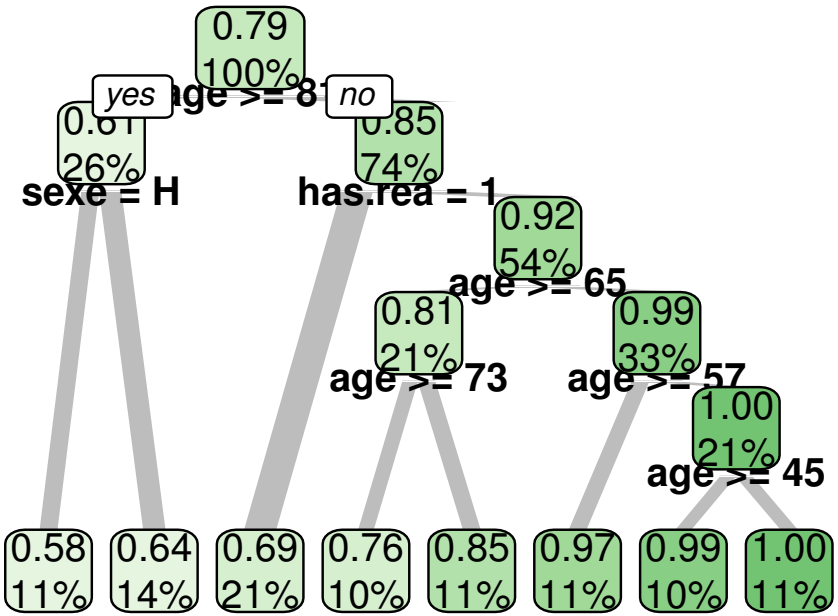




CART and learning set

Problem: CART are sensitive to perturbations in the learning set.





- Study predictions rather than structure of the tree: what is the variance associated with the sampling of the learning set?
- Bar-Hen, Gey, and Poggi, 2015: influence functions derived from robust estimation theory.
- Wager, Hastie, and Efron, 2014: variance of bagged predictors.

Robustness of CART

- Idea: quantify sensitivity through influence functions $I(\cdot)$ derived from robust estimation theory (Bar-Hen, Gey, and Poggi, 2015).
- If I is Hadamard-differentiable, then:

$$\begin{aligned}\sqrt{n}(I(F_n) - I(F)) &\simeq \sqrt{n}\int \mathrm{IC}_{I,F}(x)\mathrm{d}F_n(x) \\ &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathrm{IC}_{I,F}(X_i) \\ &\xrightarrow{\text{TCL}} \mathcal{N}\left(0, \sigma^2 = \int \mathrm{IC}^2_{I,F}(x)\mathrm{d}F(x)\right).\end{aligned}$$

- Estimation via Jackknifing:

$$\mathrm{IC}_{I,F_n}(x_i) \simeq I^*_{n,i} - I(F_n) = nI(F_n) - (n-1)I(F^{(-i)}_{n-1}),$$

where $I^*_{n,i}$ represents the n-th jackknife pseudo-value.

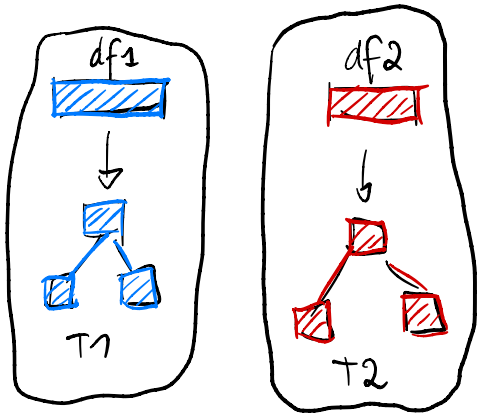
Hypothesis test for the comparison of trees

- Learning set $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, where $Y_i \in \{0, 1\}$ and $\mathbf{X}_i = (X_i^1, \ldots, X_i^p) \in \mathcal{X}^1 \times \ldots \times \mathcal{X}^p = \mathcal{X}$ with d.f. F.
- Tree $T(\mathcal{D}_n)$ generated by recursively splitting $\mathcal{X}$ along one of the $\mathcal{X}^i$ according to the minimisation of an impurity function (Gini, entropy).
- Predicted probability $p(\mathbf{x}; \mathcal{D}_n)$ of $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ for the tree $T(\mathcal{D}_n)$.
- Hypothesis test for the comparison of two trees $T(\mathcal{D}_n)$ and $T(\mathcal{D}'_m)$ on the d.f. F_n:
  - Null hypothesis $\mathcal{H}_0$: "$T(\mathcal{D}_n) \approx T(\mathcal{D}'_m)$".
  - Test statistic:

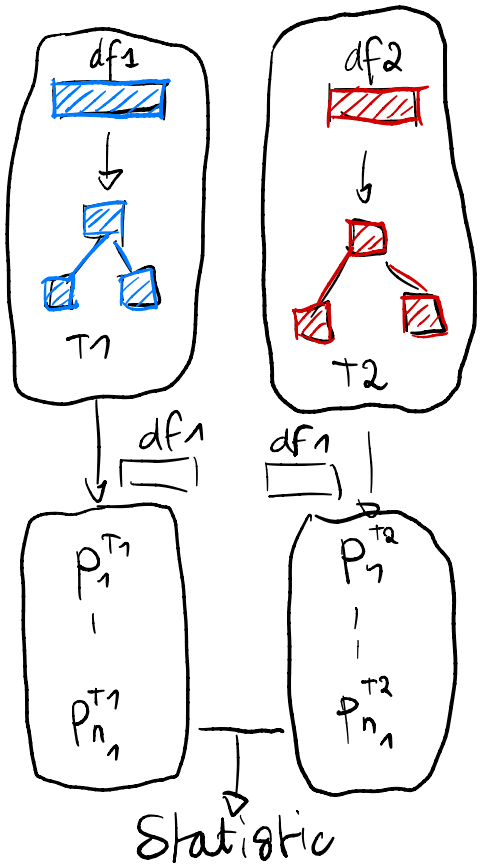
$$I(\mathcal{D}_n, \mathcal{D}'_m, F_n) = \int d\big(p(\mathbf{x}; \mathcal{D}_n), p(\mathbf{x}; \mathcal{D}'_m)\big) \mathrm{d}F_n,$$

    where $d(p, q)$ is one of $(p - q)^2$, $|p - q|$, or $-p \log q - q \log p$.
  - **Question**: What is the d.f. of $I(\mathcal{D}_n, \mathcal{D}'_m, F_n)$?

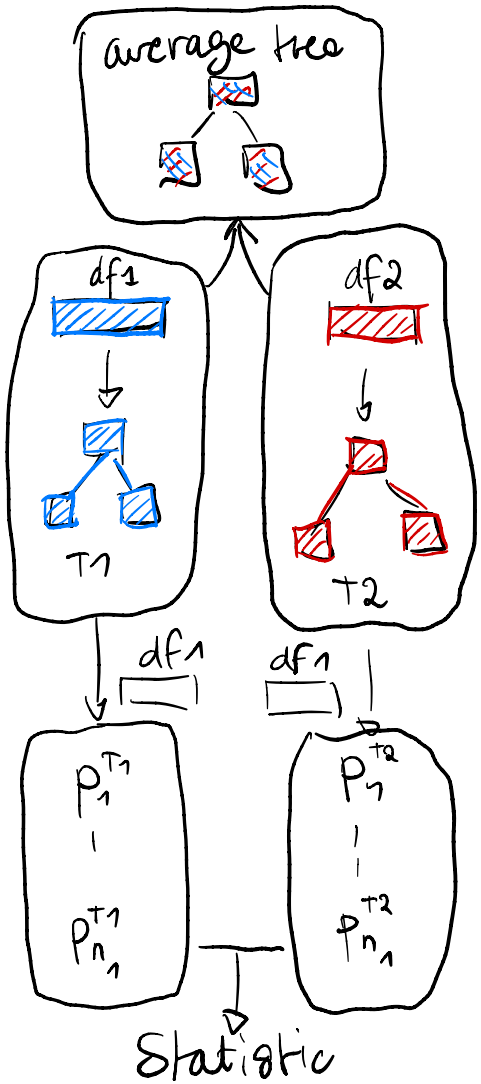
Bootstrap based hypothesis test



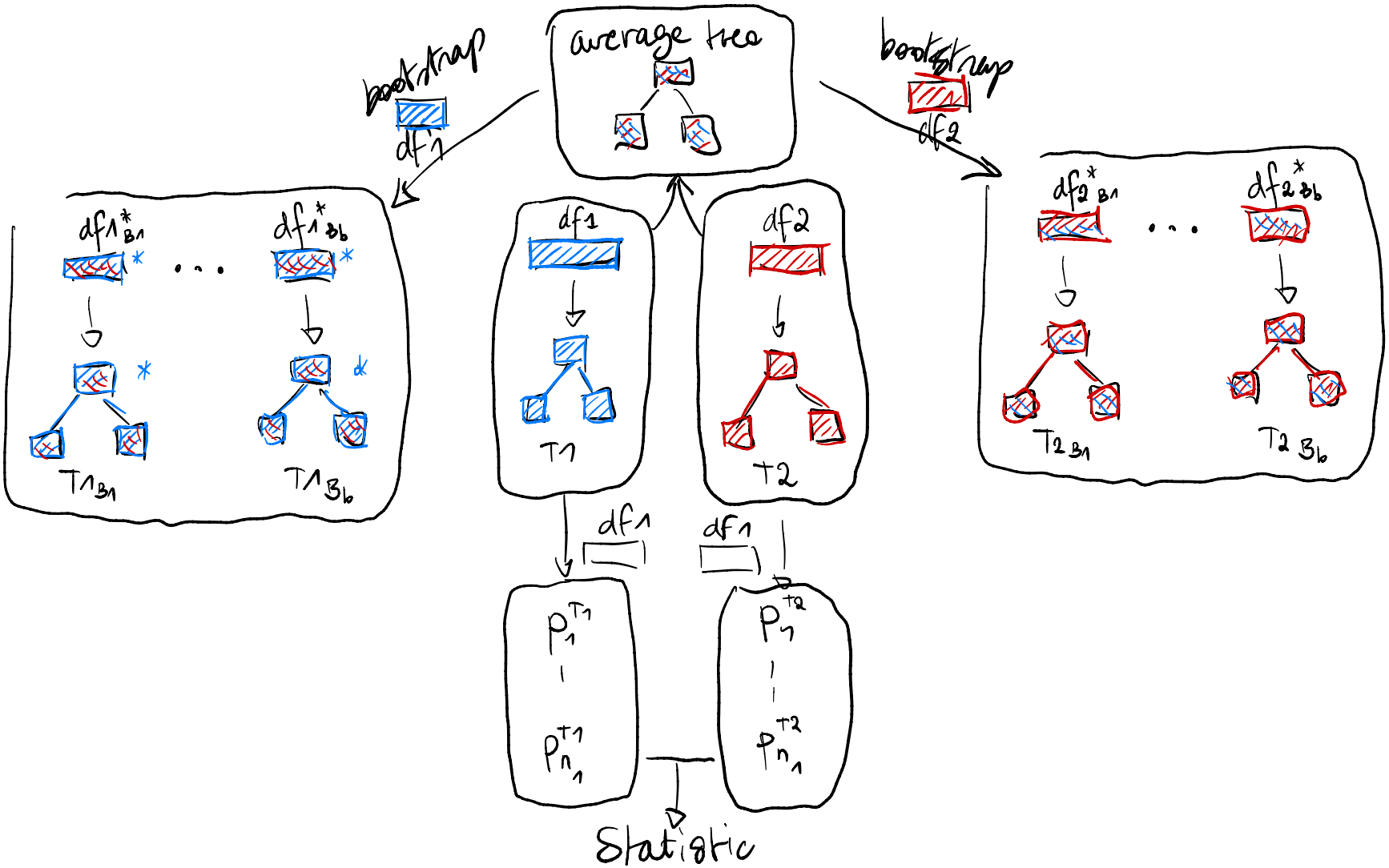
Bootstrap based hypothesis test



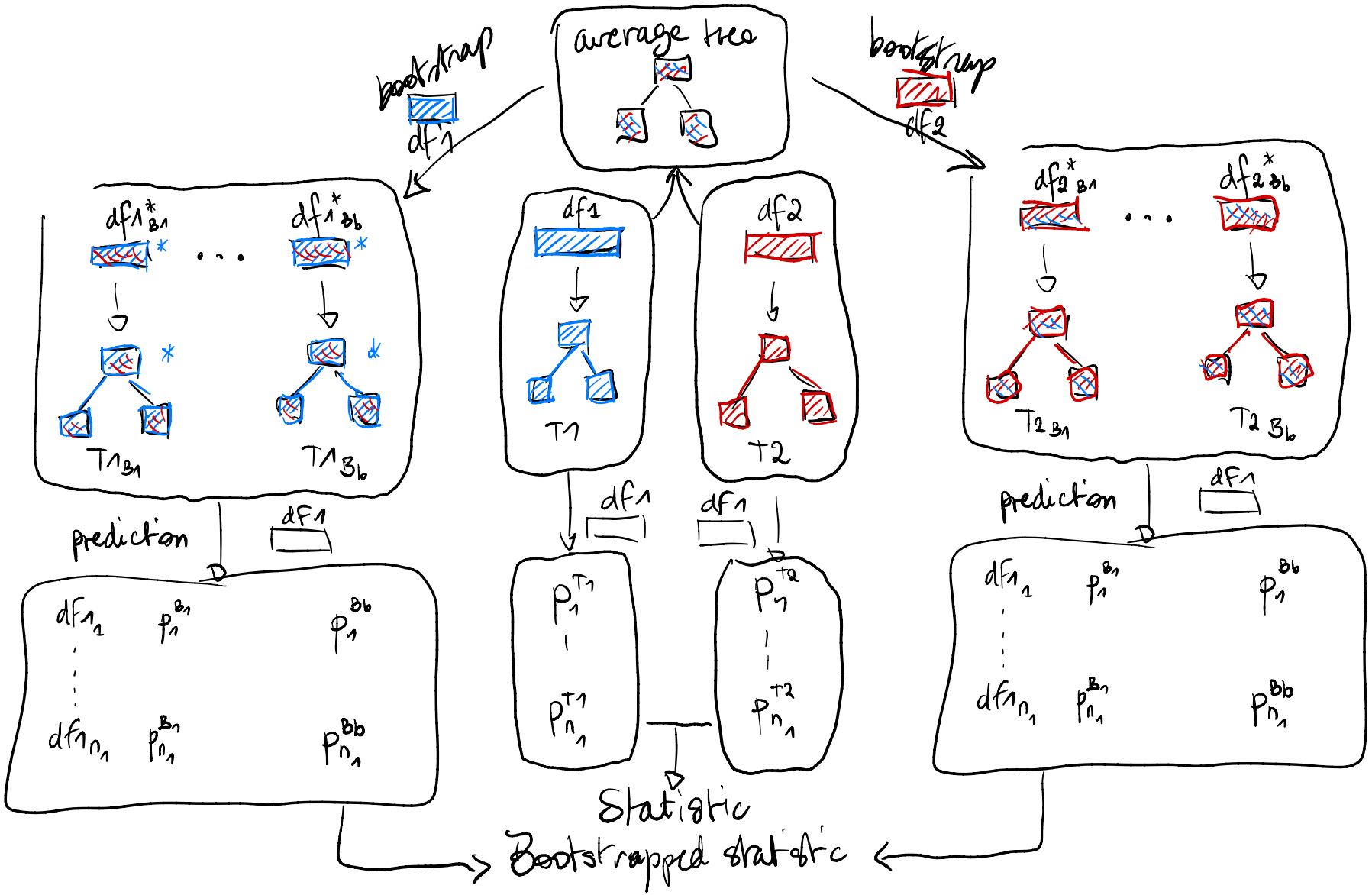
Bootstrap based hypothesis test



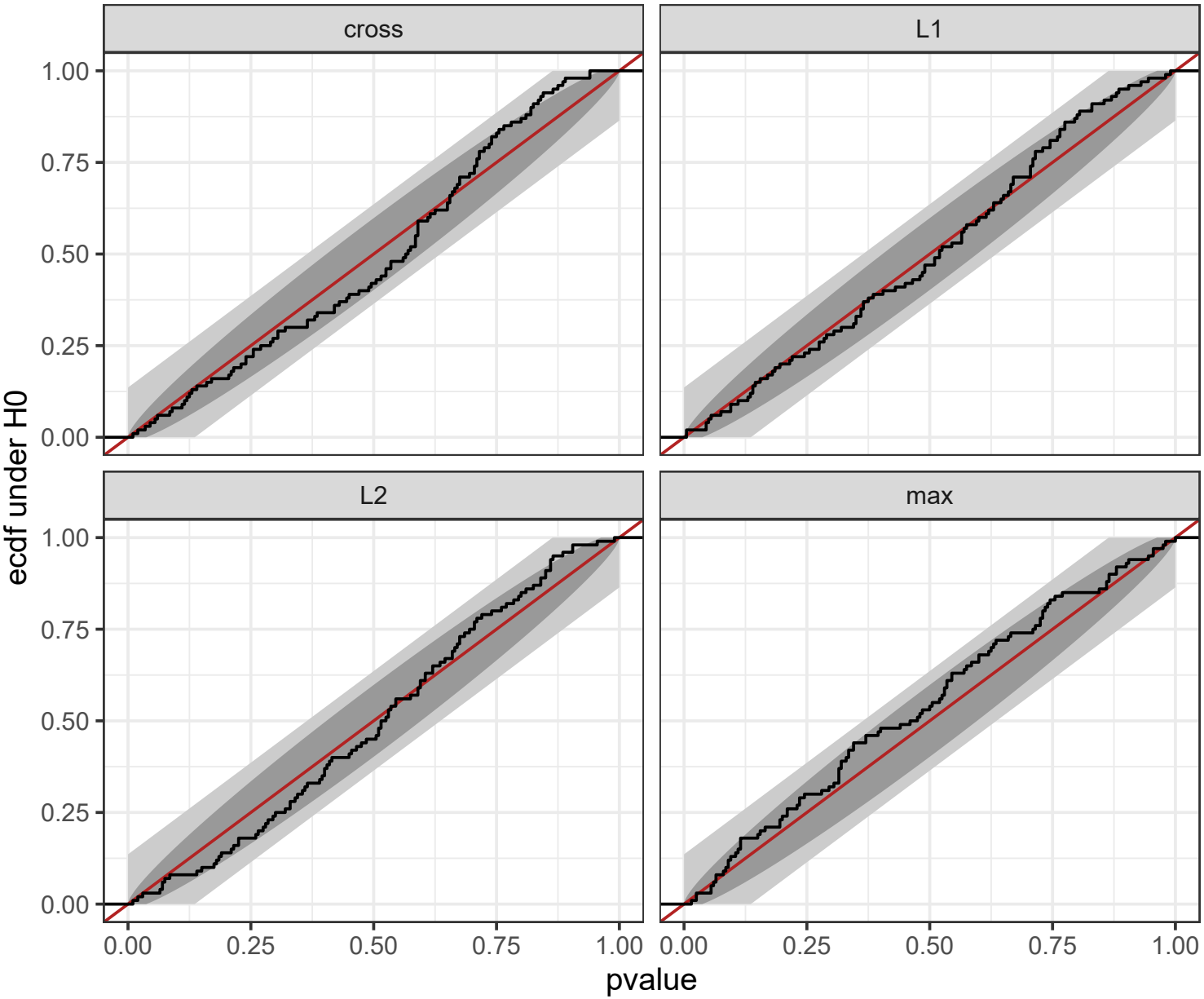
Bootstrap based hypothesis test



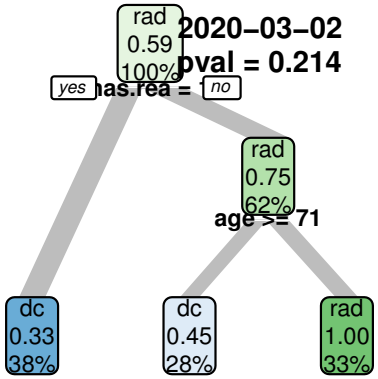
Bootstrap based hypothesis test

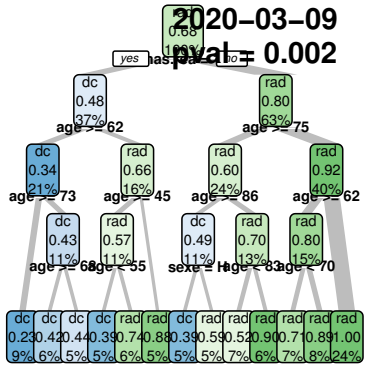


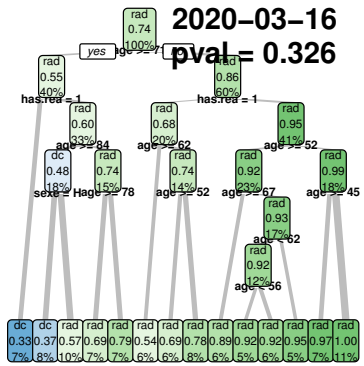
p-values for 100 simulations under $\mathcal{H}_0$

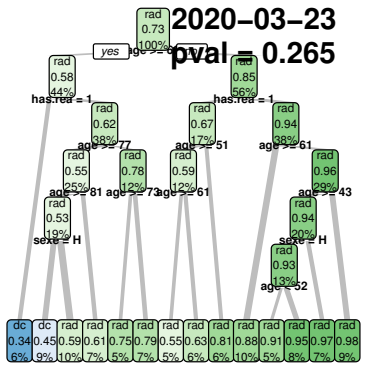


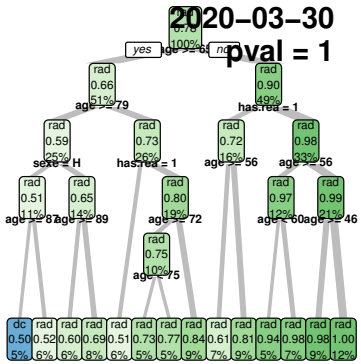
Applications to death rates during first wave

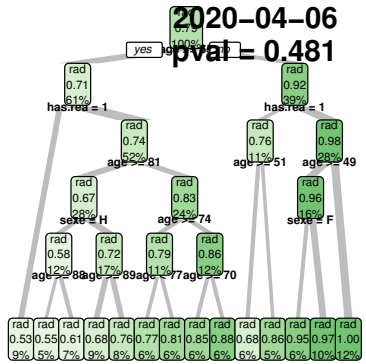


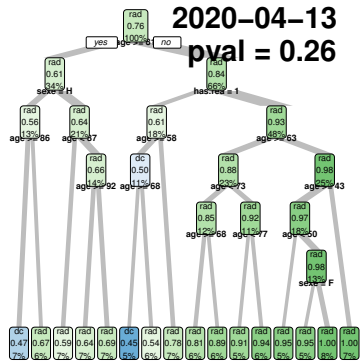


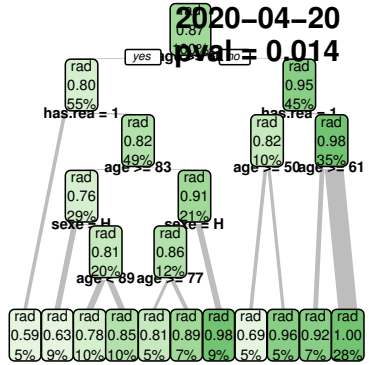


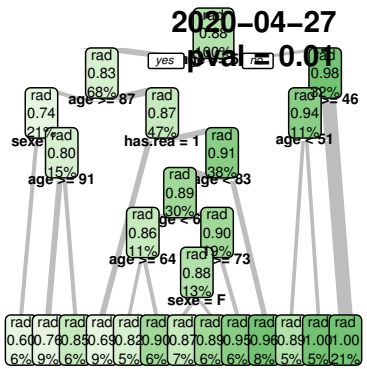


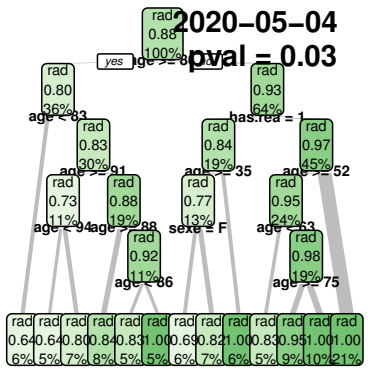


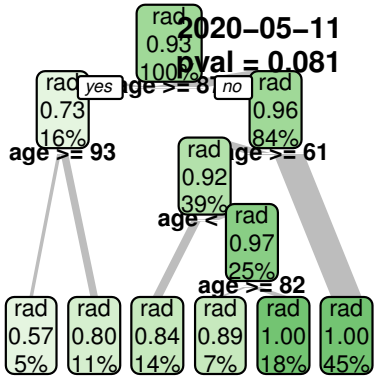


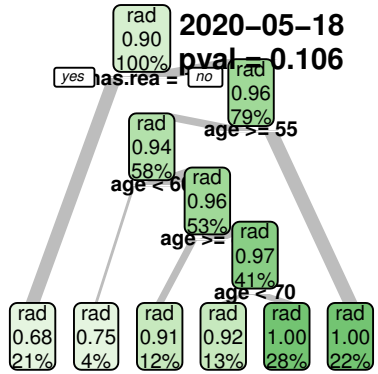












Robustness of CART

- Idea: quantify sensitivity through influence functions $I(\cdot)$ derived from robust estimation theory (Bar-Hen, Gey, and Poggi, 2015).
- If I is Hadamard-differentiable, then:

$$\begin{aligned}\sqrt{n}(I(F_n) - I(F)) &\simeq \sqrt{n}\int \mathrm{IC}_{I,F}(x)\mathrm{d}F_n(x) \\ &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathrm{IC}_{I,F}(X_i) \\ &\xrightarrow{\text{TCL}} \mathcal{N}\left(0, \sigma^2 = \int \mathrm{IC}^2_{I,F}(x)\mathrm{d}F(x)\right).\end{aligned}$$

- Estimation via Jackknifing:

$$\mathrm{IC}_{I,F_n}(x_i) \simeq I^*_{n,i} - I(F_n) = nI(F_n) - (n-1)I(F^{(-i)}_{n-1}),$$

where $I^*_{n,i}$ represents the n-th jackknife pseudo-value.

Linear statistical functionals

Consider a *linear statistical functional* $T : G \mapsto T(G) = \int \phi(x) \mathrm{d}G(x)$ with ϕ a real-valued function. Then, for the empirical d.f. F_n of F:

$$\begin{aligned}
\sqrt{n}(T(F_n) - T(F)) &= \sqrt{n} \left\{ \int \phi(x) \mathrm{d}F_n(x) - \int \phi(x) \mathrm{d}F(x) \right\} \\
&= \sqrt{n} \left\{ n^{-1} \sum \phi(X_i) - \mathbb{E}_F[\phi(X)] \right\} \\
&= \sqrt{n} \left\{ n^{-1} \sum \Big(\phi(X_i) - \mathbb{E}_F[\phi(X)] \Big) \right\} \\
&\xrightarrow{\text{CLT}} \mathcal{N}\big(0, \sigma^2 = \mathrm{Var}_F\ \phi(X)\big).
\end{aligned}$$

von Mises calculus for statistical functionals

- **von Mises differentiation** generalises this to non-linear functionals:

$$T(F_n) = T(F) + T'_F(F_n - F) + \mathrm{Rem}(F_n - F),$$

where $T'_F(\cdot - F) : G \mapsto \int \phi_F(x)\mathrm{d}G(x)$ is a linear mapping with

$$\phi_F(x) = \frac{\mathrm{d}}{\mathrm{d}t}\Big(T(F + t(\delta_x - F))\Big)\Big|_{t=0}$$

often denoted $\mathrm{IC}(x; F, T)$ the *influence curve* of T at F.

- Existence of
  - the Von Mises derivative $T'_F(\cdot - F)$
  - and convergence of $\sqrt{n}\,\mathrm{Rem}(F_n - F)$ to zero in probability,

and thus validity of the Taylor expansion, can both be ensured by the Hadamard differentiability of the functional T at F (Fernholz, 1983).

Hadamard differentiability (Fernholz, 1983)

A function $T : A \in V \to W$ is *Hadamard-differentiable* at $F \in A$ if there exists $T'_F \in L(V, W)$ such that, for any $K \subset V$ compact,

$$\lim_{t \to 0} \frac{T(F + tH) - T(F) - T'_F(tH)}{t} = 0$$

uniformly for $H \in K$. The linear function T'_F is called the Hadamard-derivative of T at F.

Perspectives

Study theoretical properties of the test

$$\sqrt{n}(I(F_n) - I(F)) \simeq \sqrt{n} \int \mathrm{IC}_{I,F}(x) \mathrm{d}F_n(x)$$

- What is the induced functional $I(\mathcal{D}, \mathcal{D}', F)$?
  - $\mathbb{L}^2$-consistency for regression trees and random forests (Scornet, Biau, and Vert, 2015):

    $$\lim_{n \to \infty} \mathbb{E}[(m_n(\mathbf{X}) - m(\mathbf{X}))^2] = 0,$$

    with $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ and $m_n(\mathbf{x})$ its prediction.
- Asymptotic properties of the bootstrap statistic $I(\mathcal{D}_n^*, \mathcal{D}_m'^*, F_n^*)$?
  - Vaart, 2000: Conditionally on $X_1, \ldots, X_n$, the sequence $\sqrt{n}(\phi(F_n^*) - \phi(F_n))$ converges in distribution to the same limit as $\sqrt{n}(\phi(F_n) - \phi(F))$, for every Hadamard-differentiable function ϕ.

Perspectives (cont'd)

Applications to more complex data from Covid-19 pandemic

- Extend methodology to include censored data.
  - Weigh observations according to the inverse of the survival function.
- Include more explanatory covariates in the learning set.
  - Biological data, comorbidities, hospital pathways, etc.
- Develop and share a R package.

Thank you for your attention.

For Further Reading I

- Bar-Hen, Avner, Servane Gey, and Jean-Michel Poggi (2015). "Influence Measures for CART Classification Trees". In: *J. Classif.* 32.1, pp. 21–45. ISSN: 0176-4268. DOI: 10.1007/s00357-015-9172-4. arXiv: 1610.08203.
- Breiman, Leo et al. (1984). *Classification and regression trees*. The Wadsworth statistics / probability series. CRC, p. 366. ISBN: 0-412-04841-8.
- Courtejoie, Noémie and Claire-Lise Dubost (2020). *Parcours hospitalier des patients atteints de la Covid-19 lors de la première vague de l'épidémie*. Tech. rep. 67. Les dossiers de la DREES.
- Fernholz, Luisa Turrin (1983). *von Mises Calculus for Statistical Functionals*. Lecture Notes in Statistics. New York: Springer-Verlag, p. 133. ISBN: 9788578110796. arXiv: arXiv:1011.1669v3.

For Further Reading II

- Scornet, Erwan, Gerard Biau, and Jean Philippe Vert (2015). "Consistency of random forests". In: *Ann. Stat.* 43.4, pp. 1716–1741. ISSN: 00905364. DOI: 10.1214/15-AOS1321. arXiv: 1405.2881.
- Vaart, A. W. van der (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, p. 460.
- Wager, Stefan, Trevor Hastie, and Bradley Efron (2014). "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife.". In: *J. Mach. Learn. Res.* 15.1, pp. 1625–1651. ISSN: 1532-4435.