What is the real prevalence of hypertension in France ?

A hierarchical Bayesian modeling approach

Édouard Chatignoux [1] Valérie Olié[1] Christophe Bonaldi [1] Amélie Gabet[1]
Clémence Grave[1] Jacques Blacher[2,3]

[1]Santé publique France

[2]Centre de diagnostic et de thérapeutique, Hôtel Dieu

[3]Université de Paris

AppliBugs, 10 décembre, 2021

Outline

Context

Hypertension (HTN)

- Permanent high blood pressure (BP) level (if not treated)
  $\rightarrow$ Systolic/Diastolic BP $\geq 140/90$ mmHg
- Leading modifiable risk factor for cardiovascular and renal diseases
- Most frequent chronic disease $\rightarrow$ majour issue in terms of resources allocation

Diagnosis of HTN

- Clinical diagnosis based on multiple BP measurements during several visits
  $\rightarrow$ Control for within subject variability
- In epidemiological studies, BP usually measured during a single visit (cost++)
  $\Rightarrow$ Biased estimates of HTN if within-person variability neglected [1]
    - $\rightarrow$ Correction made using within-person variability estimates from external studies
    - $\rightarrow$ Correction depends on the composition of the population of external studies (i.e. age and sex)

Objectives

1. Propose a method of correction that takes into account the main factors influencing BP variability : age and sex
2. Apply the method to estimate HTN prevalence in France in 2015

1. O. H. Klungel et al. "Estimating the prevalence of hypertension corrected for the effect of within-person variability in blood pressure". eng. *Journal of Clinical Epidemiology* 53.11 (nov. 2000), p. 1158-1163.

Notations and distributional assumptions

Components of BP measures

For a given sex and type of BP

Let y_{ivm} denote the m^{th} measure of blood pressure for the patient i of age $a_i = a$, during the visit v.

$$y_{ivm} = f(a) + u_i + v_{iv} + \epsilon_{ivm} \tag{1}$$

where

$$\left.\begin{array}{l} \text{- } f(a) \text{ : mean BP level for population of age } a \\ \text{- } u_i \text{ : deviation from } f(a) \text{ for individual } i \end{array}\right\} \text{ Indivudal BP level } y_i = f(a) + u_i$$

- v_{iv} : deviation from individual BP level during visit v
- ϵ_{ivm} : measurement error of the m^{th} measure during the visit v

u_i, v_{iv} and ϵ_{ivm} considered as iid gaussian random fluctuations, with variances depending on age :

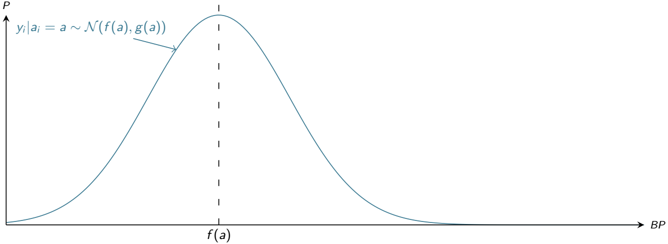
$$\begin{cases} u_i \sim \mathcal{N}(0, g(a)) \\ v_{iv} \sim \mathcal{N}(0, h(a)) \\ \epsilon_{ivm} \sim \mathcal{N}(0, l(a)) \end{cases}$$

These assumptions imply a normal distribution for y_{ivm}.

Estimator of the prevalence of hypertension I

htn = proportion of individuals with individual BP level y_i above a threshold

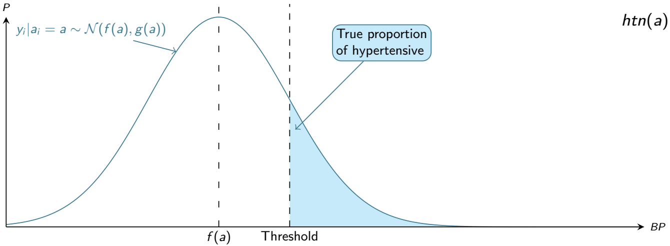
$$(y_i|a_i = a) = f(a) + u_i \sim \mathcal{N}(f(a), g(a))$$



Estimator of the prevalence of hypertension I

htn = proportion of individuals with individual BP level y_i above a threshold

$$(y_i | a_i = a) = f(a) + u_i \sim \mathcal{N}(f(a), g(a))$$



$$\begin{aligned} htn(a) &= E(y_i > T | a) \\ &= P(N(f(a), g(a)) > T) \end{aligned}$$

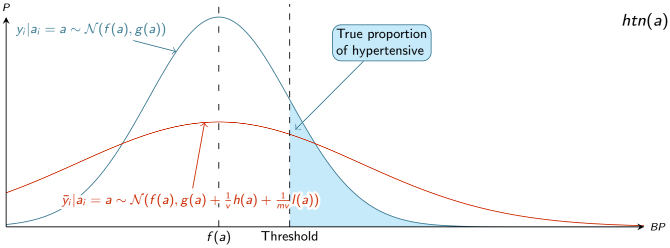
$$\Big\downarrow \textit{Empirical percentile}$$

$$\frac{1}{n(a)} \sum_{a_i = a} \mathbb{1}_{y_i > T} \qquad (2)$$

Estimator of the prevalence of hypertension I

htn = proportion of individuals with individual BP level y_i above a threshold

$$(y_i|a_i = a) = f(a) + u_i \sim \mathcal{N}(f(a), g(a))$$



$$htn(a) = E(y_i > T|a) = P(N(f(a), g(a)) > T)$$

$$\Big\downarrow \textit{Empirical percentile}$$

$$\frac{1}{n(a)} \sum_{a_i = a} \mathbb{1}_{y_i > T} \qquad (2)$$

Natural estimator for y_i

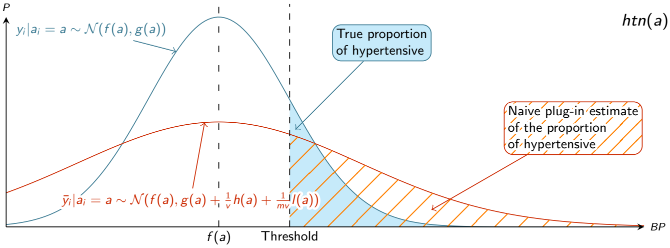
For v visits and m measures per visit :

- $(\bar{y}_i|a_i = a) = f(a) + u_i + \frac{1}{v}\sum_{k=1}^{v} v_{ik} + \frac{1}{mv}\sum_{k=1}^{v}\sum_{l=1}^{m}\epsilon_{ikl}$
- $V(\bar{y}_i|a_i = a) = g(a) + \frac{1}{v}h(a) + \frac{1}{mv}l(a) > V(y_i|a_i = a)$

Estimator of the prevalence of hypertension I

htn = proportion of individuals with individual BP level y_i above a threshold

$$(y_i | a_i = a) = f(a) + u_i \sim \mathcal{N}(f(a), g(a))$$



$$\begin{aligned} htn(a) &= E(y_i > T | a) \\ &= P(N(f(a), g(a)) > T) \end{aligned}$$

$$\Big\downarrow \textit{Empirical percentile}$$

$$\frac{1}{n(a)} \sum_{a_i = a} \mathbb{1}_{y_i > T} \qquad (2)$$

Natural estimator for y_i

For v visits and m measures per visit :

- $(\bar{y}_i | a_i = a) = f(a) + u_i + \frac{1}{v} \sum_{k=1}^{v} v_{ik} + \frac{1}{mv} \sum_{k=1}^{v} \sum_{l=1}^{m} \epsilon_{ikl}$
- $V(\bar{y}_i | a_i = a) = g(a) + \frac{1}{v} h(a) + \frac{1}{mv} l(a) > V(y_i | a_i = a)$

Plug $\bar{y}_i$ in place of y_i in (2) is biased

- Direction of the bias depends on the sign of $T - f(a)$ (i.e. positive bias if $T > f(a)$, negative otherwise)
- Magnitude increases with $\frac{1}{v} h(a) + \frac{1}{mv} l(a)$

Rescale $\bar{y}_i$ so that the resultant has the expected variance

Defining $c(a)$: correction factor for age a

$\rightarrow$ $c(a) = \sqrt{\frac{g(a)}{V(\bar{y}_i)}} = \sqrt{\frac{g(a)}{g(a)+\frac{1}{v}h(a)+\frac{1}{mv}l(a)}}$

Then

$$y_i^c = f(a) + c(a)(\bar{y}_i - f(a)) \tag{3}$$

has a gaussian distribution with mean $f(a)$ and variance $g(a)$.

$\rightarrow$ $\hat{y}_i^c$: y_i^c estimated by substituting $f(a)$ by $\frac{1}{n(a)}\sum_{a_i=a}\bar{y}_i$ in (3).

Corrected estimator

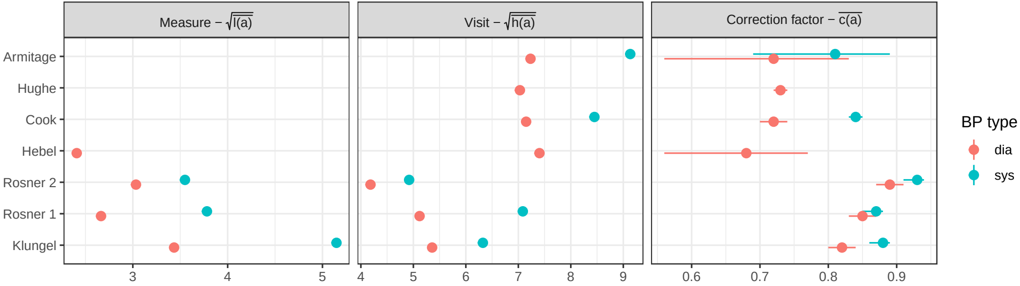
$$\hat{htn}(a) = \frac{1}{n(a)}\sum_{a_i=a}\mathbb{1}_{\hat{y}_i^c>T}$$

But we don't know $c(a)$...

Klungel - I

Use correction factor from other studies !

Study	N	Age	% women	Visits	Measure	Time v	Time m
Klungel	834	20–59	50%	2	2	1 y	5 min
Rosner 1	991	30–69	47%	3/4/5	3	1/7 days	30 sec
Rosner 2	326	0–69	37%	2/3	3	1 week	30 sec
Hebel	100	30–69	50%	2	2	3 years	5 min
Cook	2,111	16–49	100%	2	3	3 years	30 sec
Hughe	11,299	30–59	0%	4	1	1 year	–
Armitage	50	47.6	0%	4	1	1 year	–



Klungel - II

- Use a single mean correction factor (lack of detailed data)
- Corrections factors vary according to
  - The delay between visits
  - The number of measurement within visit
  - The age and sex composition of the studied population

Room for some improvement

→ Derive general shapes of the components of BP variability, by age and sex

→ Correction factor by age and sex

Outline

Estimation of $c(a)$ I

Estimation of the components of variance of y_{ivm}

$$c(a) = \sqrt{\frac{g(a)}{g(a) + \frac{1}{v}h(a) + \frac{1}{mv}l(a)}}$$

What is needed to estimate $c(a)$

Components of variability of y_{ivm} :

- $g(a)$: variability of y_i across individuals
- $h(a)$: variability of BP between visits within an individual
- $l(a)$: variability of the measures of BP within an individual during the same visit

Need data with multiple measure of BP during several visits

How to estimate the components

- ANOVA like estimates
- Hierarchical Bayesian linear models

Estimation of $c(a)$ II

Estimation of the components of variance of y_{ivm}

Hierarchical model

$y_{ivm} = f(a) + u_i + v_{iv} + \epsilon_{ivm}$ with

$$u_i \sim \mathcal{N}(0, g(a)),\ v_{iv} \sim \mathcal{N}(0, h(a)) \text{ and } \epsilon_{ivm} \sim \mathcal{N}(0, l(a))$$

- Specification of random effects (example of u_i) :
  - $\rightarrow$ Random intercept by individual $u_i^s \sim \mathcal{N}(0, 1)$
  - $\rightarrow$ Multiplied by a positive scale parameter depending on age : $\exp(g^s(a))$
  - $\Rightarrow$ $u_i = u_i^s \exp(g^s(a)) \Rightarrow V(u_i) = [\exp(g^s(a))]^2 = g(a)$
- Same for $v_{iv} = v_{iv}^s \exp(h^s(a))$ and $\epsilon_{ivm} = \epsilon_{ivm}^s \exp(l^s(a))$
- $f(a)$, $g^s(a)$, $h^s(a)$, and $l^s(a)$ estimated with penalized thin plate splines[2] :

  For a function $k(a)$: $k(a) = \alpha + \beta a + \sum_j b_j z_j(a)$
  - $\rightarrow$ $z_j(a)$: known splines basis function
  - $\rightarrow$ β and b_j parameters to be estimated
  - $\rightarrow$ Penalization of wiggliness by imposing a gaussian prior on the b_j : $b_j \overset{iid}{\sim} \mathcal{N}(0, \tau)$

---

2. Simon N. Wood. "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models". *Journal of the American Statistical Association* 99.467 (sept. 2004), p. 673-686.

Estimation of $c(a)$ III

Estimation of the components of variance of y_{ivm}

Priors

Following Gelman's[3] recommendations (default `brms` priors[4])

- Intercepts in $g^s(a)$, $h^s(a)$, and $l^s(a)$: centered Student distribution with 3 degree of freedom and a scale of 2.5
- Intercepts in $f(a)$: $\mathcal{N}(0, 10000)$
- Linear fixed effects : improper flat prior over the reals
- Standard deviations (i.e. penalties for splines) : half student-t prior with 3 degrees of freedom and a scale of 2.5.

Estimation

- Hamiltonian Monte Carlo with Stan software[5]
- 4 chains with 6 000 iteration (5 000 burn-in)
- Numerical computations performed on the S-CAPAD/DANTE platform, IPGP, France

3. Andrew Gelman. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". *Bayesian Analysis* 1.3 (sept. 2006), p. 515-534.
4. Paul-Christian Bürkner. "brms : An R Package for Bayesian Multilevel Models Using Stan". *Journal of Statistical Software* 80.1 (2017), p. 1-28.
5. Stan Development Team. *stan Modeling Language Users Guide and Reference Manual, 2.27.* 2021.

Data from NHANESIII study - 1988-1994

- 18,825 adults from general US population ($\geq$ 17 y.o.)
- 2 or 3 (n=2,174) visits, 2 BP measurements per visit
- First visit to a mobile examination center
- Median duration between subsequent consecutive of 17 days (minimum 1 day, maximum 48 days)

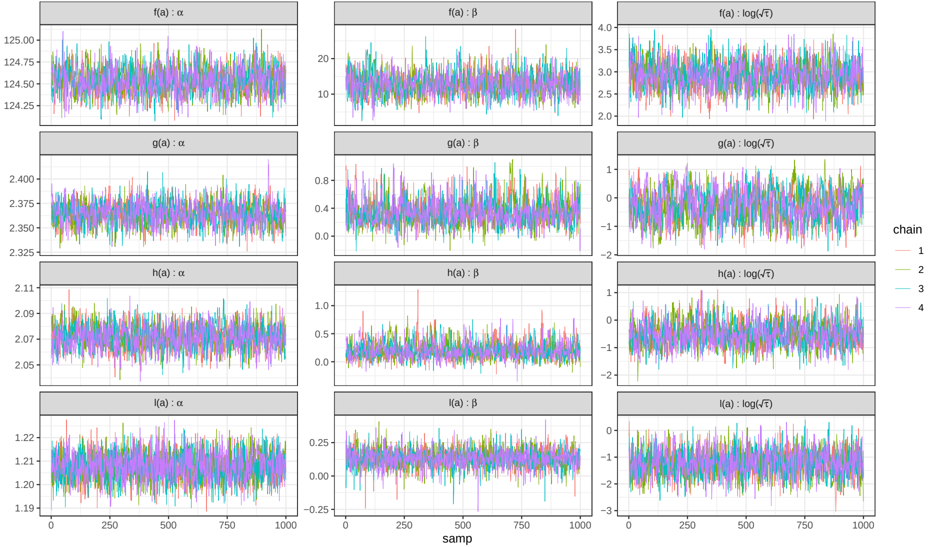
Hierarchical model estimated separately by :

- Sex
- Type of blood pressure (i.e. systolic and diastolic)
- Patients taking or not anti-hypertensive treatments

Outline

Results - convergence

Traceplots of model parameters for systolic blood pressure in men - untreated patients



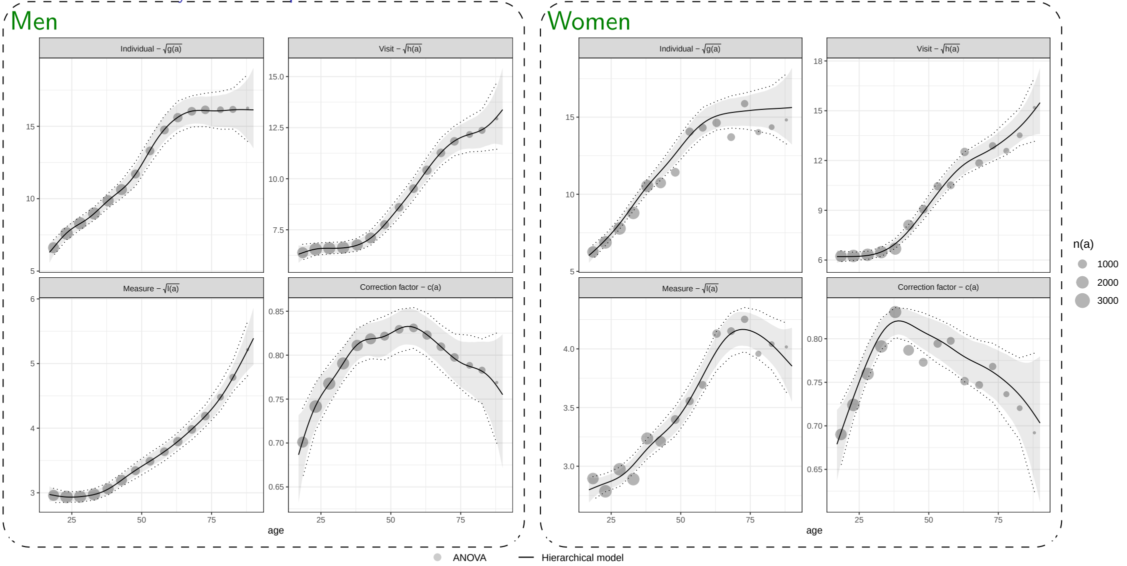
Results - convergence

$\hat{R}$ and ESS - untreated patients

BP type	gp	par	Men $\hat{R}$	Men ess (bulk)	Men ess (tail)	Women $\hat{R}$	Women ess (bulk)	Women ess (tail)
Diastolic	f(a)	α	1.00	1,399.30	2,160.29	1.00	2,121.48	2,380.68
		β	1.00	1,304.41	1,681.94	1.00	2,379.13	2,917.31
		$\sqrt{\tau}$	1.00	1,014.70	1,407.12	1.00	1,575.97	2,183.74
	l(a)	α	1.00	2,214.12	2,910.38	1.00	1,939.27	2,856.58
		β	1.00	1,563.06	2,166.84	1.00	2,491.48	2,484.32
		$\sqrt{\tau}$	1.00	1,214.76	2,023.57	1.00	1,133.40	2,119.20
	g(a)	α	1.01	662.16	1,325.92	1.00	955.40	1,959.27
		β	1.00	980.40	922.60	1.00	1,237.86	1,822.75
		$\sqrt{\tau}$	1.00	660.84	1,196.00	1.00	846.64	1,335.66
	h(a)	α	1.00	940.02	1,716.07	1.00	1,096.93	1,900.19
		β	1.00	1,317.91	1,901.86	1.00	748.18	1,513.16
		$\sqrt{\tau}$	1.00	927.64	1,808.28	1.01	673.37	1,029.30
Systolic	f(a)	α	1.00	1,045.55	1,952.33	1.00	2,082.74	2,349.82
		β	1.00	1,345.91	1,894.80	1.00	1,997.69	2,079.77
		$\sqrt{\tau}$	1.01	1,107.39	2,051.83	1.00	1,786.75	2,272.82
	l(a)	α	1.00	2,183.92	3,091.98	1.00	1,357.95	2,692.41
		β	1.00	2,220.05	2,775.54	1.00	2,684.52	2,776.89
		$\sqrt{\tau}$	1.00	1,679.53	2,336.00	1.00	2,211.08	2,568.53
	g(a)	α	1.00	1,084.71	1,921.31	1.00	1,596.61	2,204.35
		β	1.01	899.73	1,694.29	1.00	1,823.53	2,636.01
		$\sqrt{\tau}$	1.01	600.28	1,267.82	1.00	1,907.28	2,238.26
	h(a)	α	1.00	1,072.68	2,223.07	1.00	1,206.96	2,391.47
		β	1.00	1,171.63	1,426.56	1.00	1,958.17	2,766.99
		$\sqrt{\tau}$	1.00	908.50	1,736.86	1.00	2,094.04	2,719.96

Results (untreated patients)

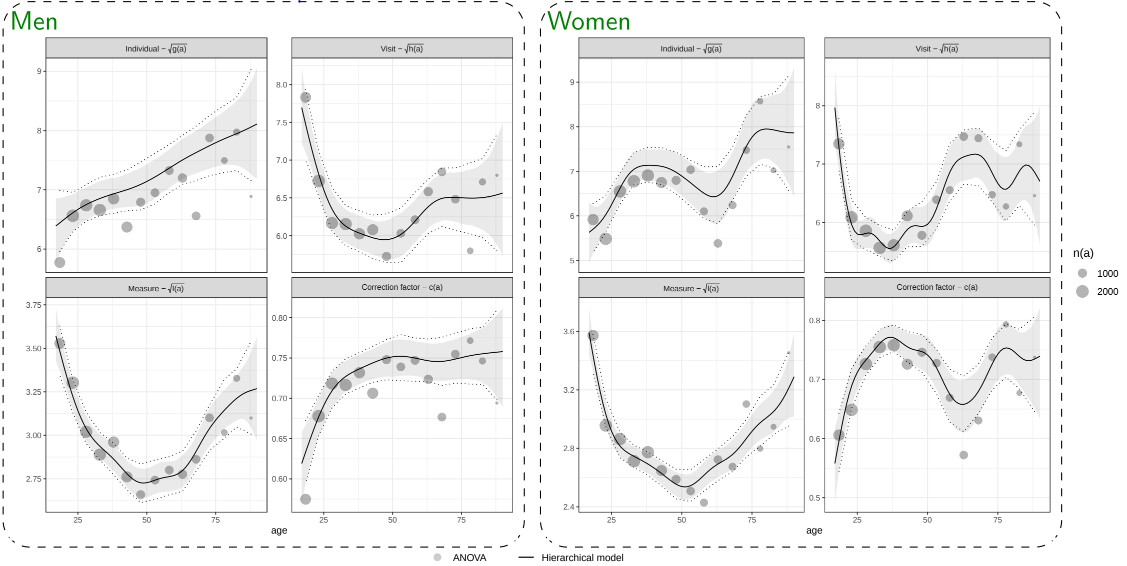
Components of variance : Systolic blood pressure



Note : $c(a)$ calculated for $v = 1$ and $m = 2$

Results (untreated patients)

Components of variance : Diastolic blood pressure



Note : $c(a)$ calculated for $v = 1$ and $m = 2$

ESTEBAN study

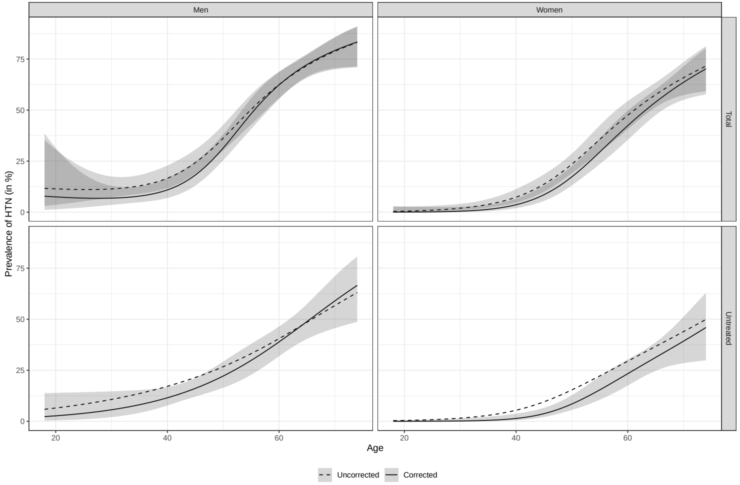
- Cross-sectional study (2014-2016)
- $\sim$ 2,000 individuals 18 to 74 y.o.
- 2 BP measures during a single visit

Estimation

1. For each post-sample of $c(a)$
   - Correct individual BP $\rightarrow y_i^c$
   - Estimate HTN (using sampling weights)
     - $y_i^c >$ threshold OR
     - Patient treated for HTN
2. Combine post-sample's estimates
   - Variance = mean variance of post-samples + variance across post-sample's estimates

HTN prevalence in France

Prevalence by age



Overall

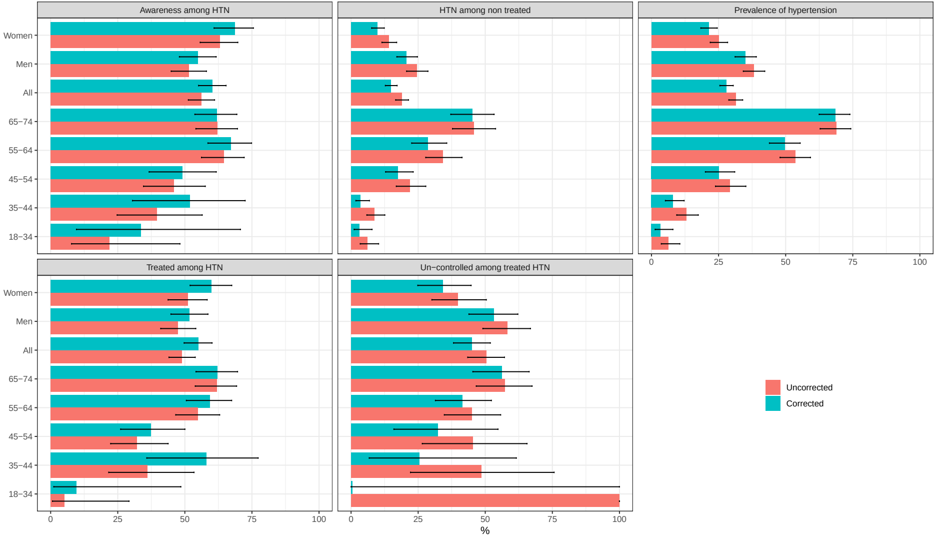
	Un-corrected	Corrected
Men	38.1[34.2 ;42.2]	35.0[31.2 ;39.1]
Women	25.0[21.9 ;28.4]	21.3[18.4 ;24.5]
All	31.3[28.8 ;34.0]	27.9[25.5 ;30.5]

- Larger differences in women than in men
- Larger differences in young than in elderly

→ 12.7 instead of 14.3 millions of cases for the 18-74

HTN prevalence in France

Effect of correction in sub-pops



Outline

Discussion

Method

- Control for differences in age and sex composition of study (e.g. ESTEBAN) vs reference study (e.g. NHANESIII)
  - Main factors driving variability of BP
  - Easy to apply to subpop
- Main hypothesis : $c(a)$ estimated from external data applies to study
  - $\rightarrow$ Less restrictive than equality of variances
  - $\rightarrow$ Compatibility between populations/study protocol ?
  - $\rightarrow$ In our case, the variability of y_{ivm} observed in ESTEBAN $\simeq$ predicted from NHANESIII component of variance
- Other factors influencing BP variability not accounted for

Hierarchical modeling

- Gaussian assumption
- No correlation between components of variance

Discussion

Results

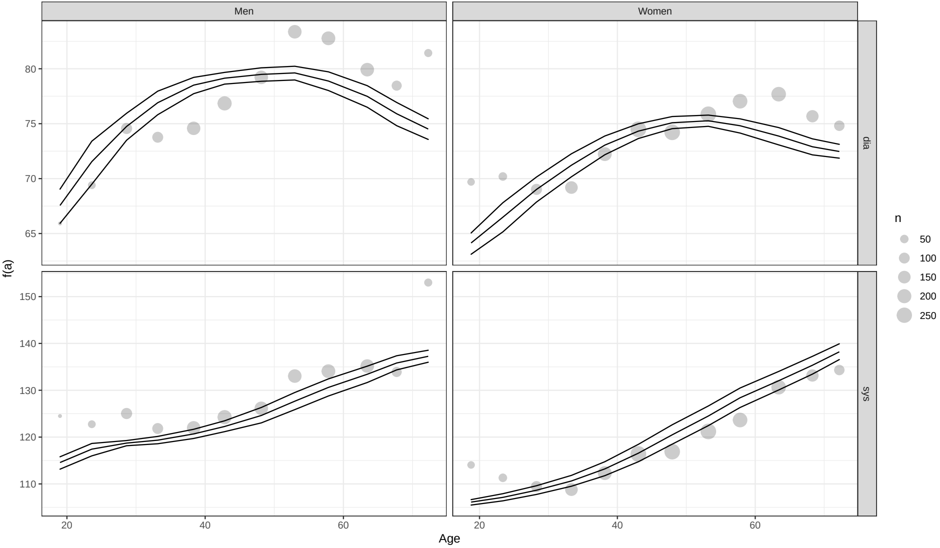
- Substantial variations of $c(a)$ with age and sex
- Modest to substantial correction of HTN
- Within CI bands

R package

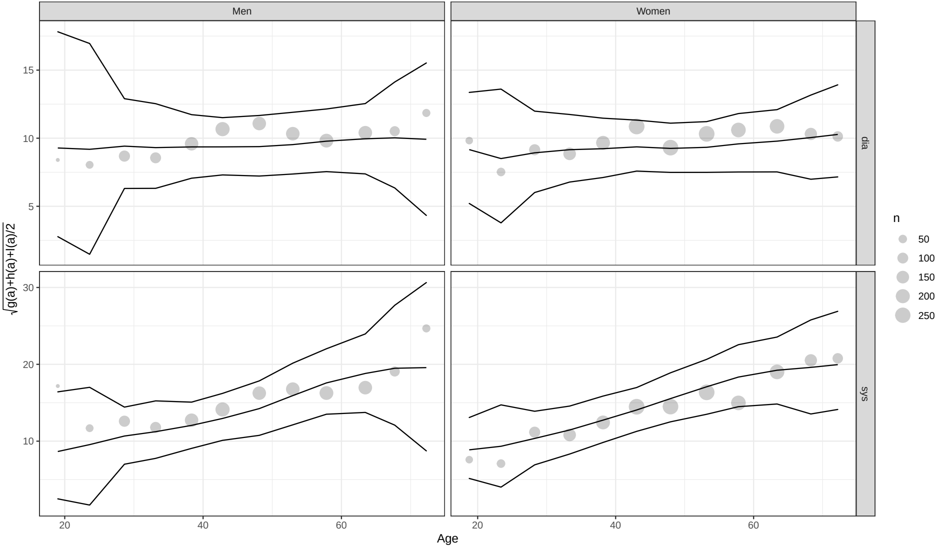
Method of correction disseminated in a R package available at
`https://github.com/echatignoux/BPpack`

Appendix

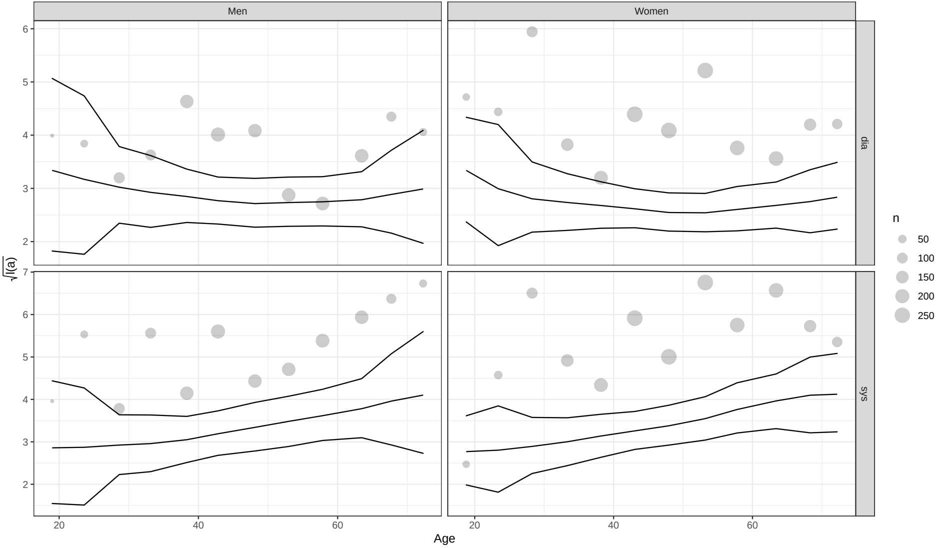
Mean BP levels in Esteban vs Mean BP levels predicted from NHANES



Total variance in Esteban vs total variance predicted from NHANES



Measurement error in Esteban vs measurement error predicted from NHANES



ANOVA derivation of variance components

If m measures of BP are realized during v visits, analytical estimator of h, g and l can be derived using an ANOVA approach.

If we note $\bar{y}_{iv}$ the mean of BP measures for individual i during visit v, then $V(y_{ivm} - \bar{y}_{iv}) = \frac{m-1}{m} l(a)$, leading to

$$\hat{l}(a) = \frac{m}{m-1} V(y_{ivm} - \bar{y}_{iv} | a_i = a)$$

Similarly, $V(y_{ivm} - \bar{y}_i) = \frac{v-1}{v} h(a) + \frac{mv-1}{mv} l(a)$, so $h(a)$ may be estimated by

$$\hat{h}(a) = \frac{v}{v-1} V(y_{ivm} - \bar{y}_{iv} | a_i = a) - \frac{m-1}{m(v-1)} \hat{l}(a)$$

An estimator for $g(a)$ derives from the expressions above :

$$\hat{g}(a) = V(y_{ivm} | a_i = a) - \hat{h}(a) - \hat{l}(a)$$