FNSNF Swiss National Science Foundation
The National Centres of Competence in Research (NCCR)
are a research instrument
of the Swiss National Science Foundation
UNIVERSITÉ DE GENÈVE
UNIL | Université de Lausanne

A Data Mining Approach to Longitudinal Risk Assessment in Cognitive Epidemiology

Stephen Aichele & Paolo Ghisletta

LIVES
Swiss National Centre of Competence in Research

Random Forest Analysis

Algorithmic emphasis (computational methodology with no overarching parametric model specified a priori)

Exploratory in the sense that it performs an exhaustive search for the "best predictors" of a given outcome

Attractive features:

- Can account for complex relations between variables (e.g., truly non-linear relations, higher-order interactions)
- Can accommodate large numbers of predictor variables (even given relatively few observations)
- Robust against multicollinearity and selection bias (with respect both to variables and to observations)

RFA is an Extension of CART

Classification and Regression Trees (CART)

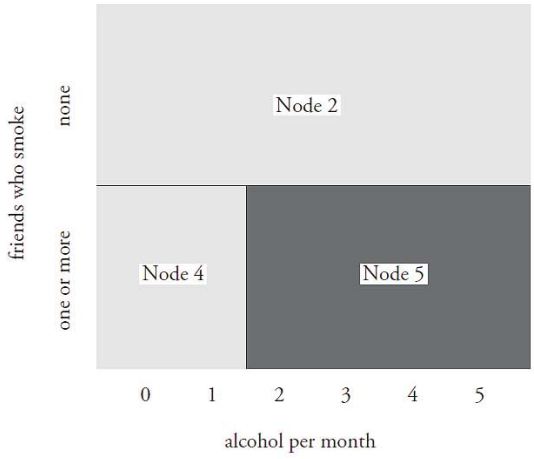
Breimen et al. (1984); Quinlan (1986, 1993)

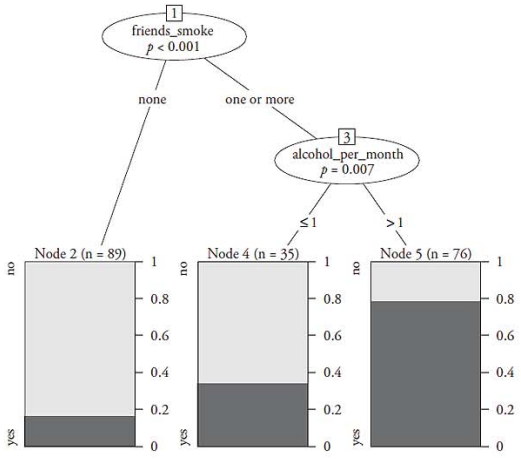
CART algorithm uses “recursive partitioning”: A sample is split along a given variable into sub-samples, sub-samples are again split, and so on

This forms an upside-down “tree” of (typically) bifurcating “nodes”, such that a given path from the root node to an end node (or leaf) represents a series of predictor/cut-point decisions that maximally differentiate the sub-samples at each step with respect to the outcome of interest.

Recursive Partitioning

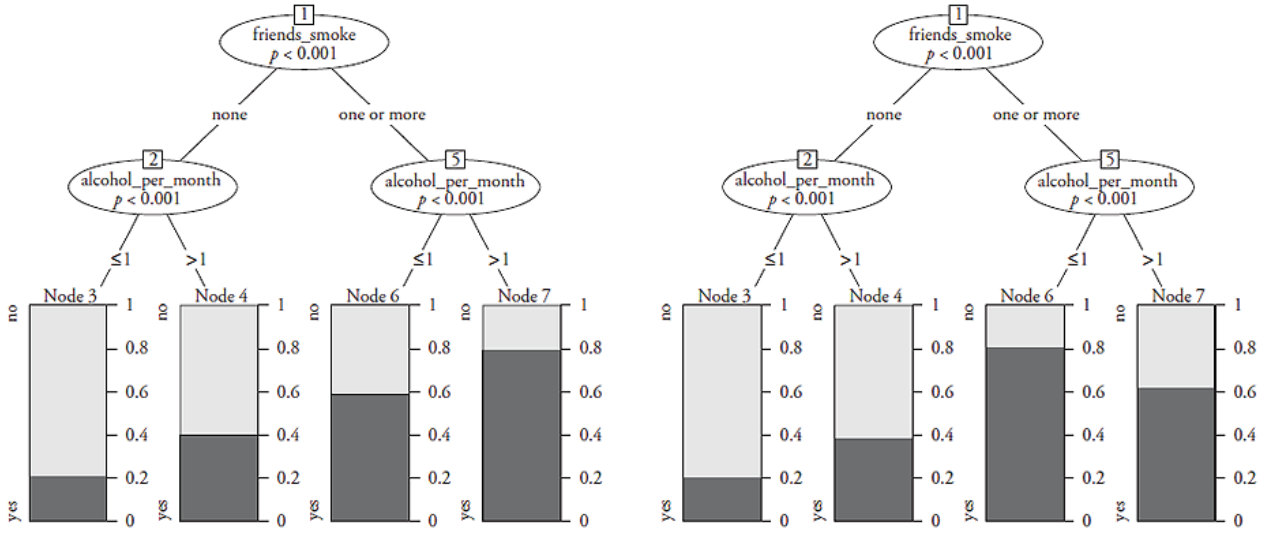
The range of predictor relations includes all combinations of rectangular partitions that can be derived by recursive splitting (including multiple splits in the same variable). Thus, nonlinear and even non-monotone association rules are accounted for.





Images adapted from Carolin Strobl, in *The Oxford Handbook of Quantitative Methods II*

Illustration of Regression Trees



Regression trees with two main effects (left) and interactions (right)

Image adapted from Carolin Strobl, in *The Oxford Handbook of Quantitative Methods II* (p. 689)

Recursive Partitioning

However, if the relationship was truly linear, the approximation given by CART would be inferior to that provided by standard regression. But if the linear relation involved a higher-order term (e.g., *x*^3), CART would at least approximate it “by default”, whereas failure to include the term in the linear model would reduce model fit.

Variable Selection Algorithm

At each node of the tree (i.e., within each sample/sub-sample), need to determine:

by which predictor

and at which corresponding cut-point

observations can be subdivided to maximally determine (i.e., predict) differences in the outcome of interest.

Usually involves minimization of a loss function (e.g., Shannon or GINI entropy for categorical outcomes, mean squared error for continuous outcomes)

Predictor/Cut-Point Selection Algorithm

Classical Method:

At each node (within each nested sub-sample)
- For each cut-point in each predictor
  - Split node & calculate difference in predictor/outcome association between each of the daughter nodes
- Select variable/cut-point that maximizes this difference
- Recurse within child nodes until stopping criterion (e.g., min. *n*)

Key Shortcomings of this approach (Strobl):

1. Continuous variables and categorical variables with more categories (and hence possible cut-points) are favored (more simultaneous comparisons)
2. For some measures of association, variables with more missingness may be preferred

Predictor/Cut-Point Selection Algorithm

A better (2-step) approach:

At each node (within each nested sub-sample)

(Step 1)

For each predictor

Compare change in strength of association when values of predictor are randomly permuted (while values of covariates remain fixed) => permutation accuracy test

Do any of the predictors have significant association with outcome?

If no: stop

If yes: pick predictor with strongest outcome

(Step 2)

For strongest predictor, determine best cut-point

Recurse until stopping condition is met (no longer significant prediction, minimum sample size reached) or grow a full tree and then "prune"it back

Limitations of Single-Tree Method

Spurious variable selection

When variables are highly correlated, the decision to select one over the other becomes increasingly a matter of chance. Spurious variable selection closer to the root node will, on average, more strongly bias the resulting tree. Thus, the structure of a given tree is inherently "unstable"

Model over-fit

When a tree is fit to the entire pool of observations, model overfit (sampling both signal and noise) is possible (i.e., no built-in mechanism for cross-validation). The resulting tree representation may not hold for new samples.

Random Forest Analysis

Ensemble methodology: Instead of 1 tree, "grow" a forest (numerous trees) & aggregate results across trees

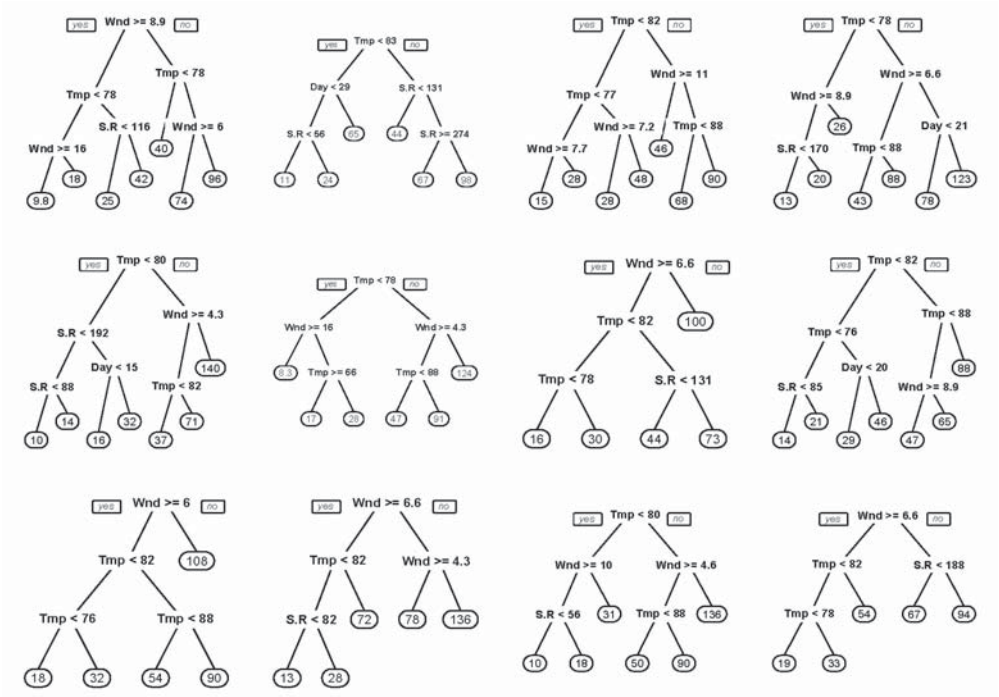
Randomly draw sub-samples

(1) of observations used in each tree (i.e., bootstrap aggregation or "bagging")

(2) of predictor variables at each node of a given tree (typically a small subset (*mtry*)

Pool estimates of predictors' variable importance across trees

Random Forest Analysis (only 12 trees)



Variable Importance (VIMP)

There are different measures of variable importance (VIMP). Some analysts prefer "permutation accuracy": Calculate loss function (e.g., mean squared error) before and after randomly permuting values of the given predictor, then take the difference. Average these (weighted) estimates of permutation accuracy across all trees in the forest.

Note that the absolute scale of VIMP can change on a forest-to-forest basis. So predictors' VIMP estimates should be reported as relative importance (percentage of maximum VIMP, or as rank order)

Random Forest Analysis

Addresses problems of model overfit (to samples or to variables) and instability

Variables selected at nodes within a given tree are based on "in bag" observations, but variable importance can later be estimated using "out of bag" observations (cross-validation)

The ensemble methodology gives each variable a more fair chance to appear in different configurations (i.e., a check against multicollinearity and spurious variable selection)

RFA Issues (according to Strobl)

Variable selection bias can still be a problem if different variable types are used and if the tree-level algorithm favors variables with more potential cut-points

Variable selection bias can also occur when using bootstrapped samples of observations for each tree (e.g., sampling with replacement) rather than sub-samples (sampling without replacement)

RFA Issues (according to Louppe)

The main problems aren't related directly to VIMP or variable type, rather:

Growing trees with too few observations in the "leaves" (final sub-samples) will lead to increased noise in estimates of VIMP (in RFA, trees are often grown very deep)

Variable sub-sampling at nodes can introduce another form of bias (e.g., because variable comparison at a node will influence variables selected subsequently)

RFA Issues (according to Louppe)

Therefore, a different strategy:

- Create a forest of "completely random" trees in a single step by random selection of a single variable (*mtry* = 1) and random cut-point at each node of a tree
- Limit depth of trees (to limit prediction error in final nodes/leaves) – optimal will depend on sample size
- Calculate each variable's VIMP across the entire forest (rather than as an average of tree-specific VIMP)

This is currently implemented in the Python machine learning toolbox: *scikit-learn*

RFA and Missing Data

One of the selling-points of RFA is that it can account for complex relations (non-linear associations, higher-order interactions) between variables.

But RFA requires complete data, and by default most imputation methods only consider linear main effects to identify candidate values. Single imputation of missing values using random forest estimation is built in to some RFA software packages, and R package 'mice' no provides RFA with multiple imputation, but it is still not clear how best to pool VIMP estimates across imputations.

RFA – Other Concerns

Interpretation of the functional form underlying a forest generally not possible – trees are not nested, and there is no such thing as an "average tree" (i.e., black box method)

Authors of RFA software don't always agree on preferred implementation and have different ideas as to sources of bias. RFA software code, even when open source, can be difficult to comprehend as it is wrapped around "lower level" languages (such as C) in order to optimize performance

RFA is Useful

Despite these limitations…

RFA can be very effective for identifying key variables based on their relative strength of association – and used as a precursor to analyses based on structural (theoretical) models of those associations

Application in Cognitive Epidemiology

Cognition & Survival

Early scientific interest in the association between cognitive ability & survival :

- Maller (1933)
- Kleemeier (1962) , Riegel and Riegel (1972), Siegler (1975)

Childhood IQ predicts mortality risk up to age 70.

- E.g., Whalley & Deary (2001)

Age-related cognitive decline is linked to increased mortality risk.

- E.g., Rabbitt et al. (2002)

Relation persists even after adjusting for demographic variables (gender, education level) and dementia diagnosis.

Manchester & Newcastle Longitudinal Study of Cognition

20+ year study of 6203 individuals

Age at initial testing: *m* ≅ 66 years, range = 41-96 years

Multiple measures of cognitive performance: crystallized abilities, fluid abilities, visuospatial memory, verbal memory, processing speed - *repeatedly assessed* (up to 4 times) over a 12 year period

Lifestyle, behavioral, and health indices (also repeatedly assessed)

Goals of the Study

Determine which cognitive abilities are the most sensitive predictors of mortality risk?

Determine whether life span *decrements* in cognitive performance predict mortality risk after adjusting for *baseline* ability?

Compare the relative influence of multiple cognitive, demographic, lifestyle, behavioral, and health variables as concurrent predictors of mortality risk

Preliminary Analyses

Multiple Imputation (to handle missing data)

Structural Factor Models (Aggregation across Measures)

- cognitive abilities
- functional health

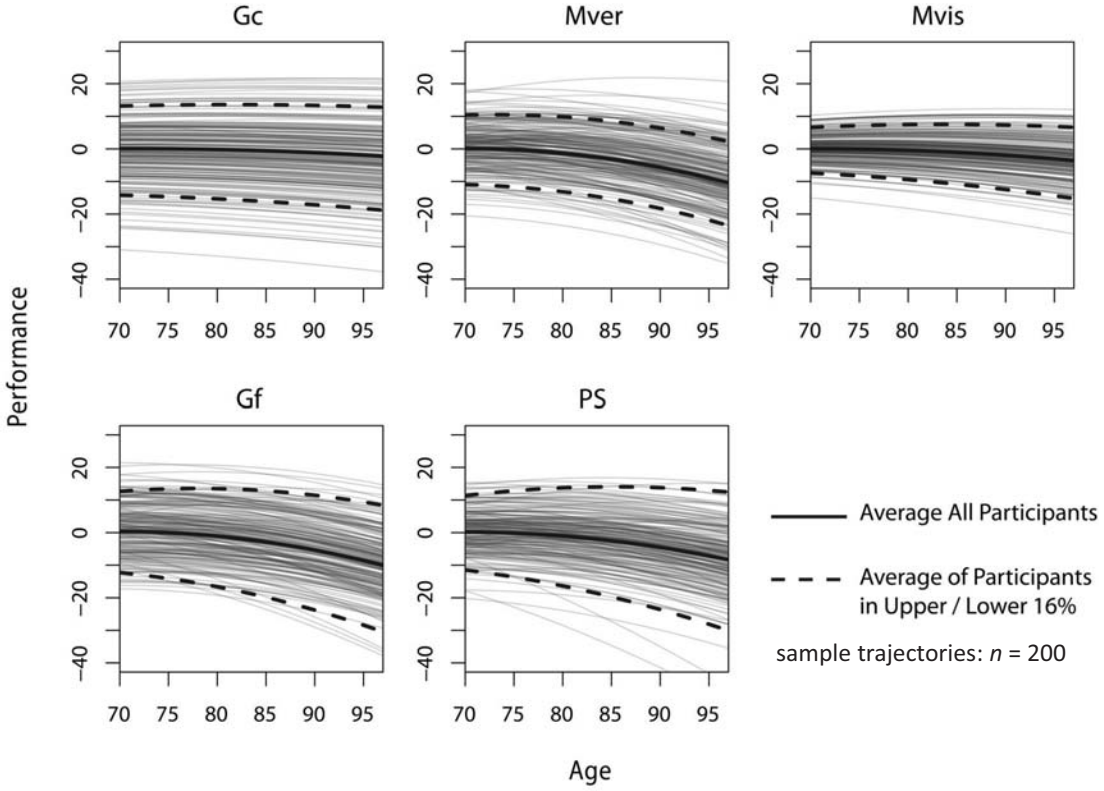
Multi-Level Growth Models (Aggregation across Time)

- baseline levels and life span changes in cognitive, health, and lifestyle indices

Longitudinal Factor Analyses of Cognitive Abilities

Domain	Cognitive Task	λ
Gc	Raven Mill Hill Vocab. A	.90
	Raven Mill Hill Vocab. B	.90
	WAIS-R Vocabulary	.84
Gf	Heim Intelligence Test 1	.89
	Heim Intelligence Test 2	.90
	Cattell Culture Fair Test	.83
Mver	Free Verbal Recall	.75
	Cumulative Verbal Recall	.86
	Delayed Verbal Recall	.70
Mvis	Picture Recognition	.44
	Memory Objects	.69
	Shape + Spatial Locations	.50
PS	Visual Search	.73
	Alphabet Coding Task	.85
	Semantic Reasoning	.72

Estimated Life Span Changes in Cognitive Performance



Comparative influence of cognitive, demographic, lifestyle, & health variables on mortality risk

Socio-Demographic (8)
Age at Study Induction
Gender
City of Residence
Study Cohort
Occupational Class
Marital Status
Persons in the Home
Children

Tobacco and Alcohol Use (5)
Smoker?
Years Smoking
Drinker?
Years Drinking Alcohol
Daily Alcohol Consumption

Daily Life Attributes (16)
Subjective Health
Prescribed Medications
Sleep (hr./night)
Wake-ups (tot/night)
Number of Hobbies
Difficulty Doing Housework
Impaired Physical Mobility
Leisure Activity
Casual Contacts (tot/wk)
Short Conversations (tot/wk)
Long Conversations (tot/wk)

Cognitive Abilities (10)
Crystallized Intelligence
Fluid Intelligence
Verbal Memory
Visual Memory
Processing Speed

CMI Medical Symptoms (26)
A: Eyes/Ears
B: Nose/Throat/Respiration
C: Cardiovascular
D1: Teeth
D2: Gastrointestinal/Liver
E: Musculoskeletal
F: Skin
G: Nervous System
H: Reproductive/Urinary
I: Fatigue
J: Frequency Illness
K: Miscellaneous
L: Addiction
M: Inadequacy
N: Depression
O: Anxiety
P: Sensitivity
Q: Anger
R: Tension
Total symptoms (A–L , M–R)

Statistical Methodology

Split observations into two sub-samples

Random Forest Survival Analysis (RFSA) [n_1]

Algorithmic approach: Uses repeated randomized sampling of observations and predictors within increasingly smaller nested subsamples

Estimates of predictor influence are implicitly adjusted for higher-order interactions and insulated from bias due to order effects, model over-fit, and multicollinearity

Cox PH model of most important predictors [n_2]

Effect size estimation using known statistical distribution

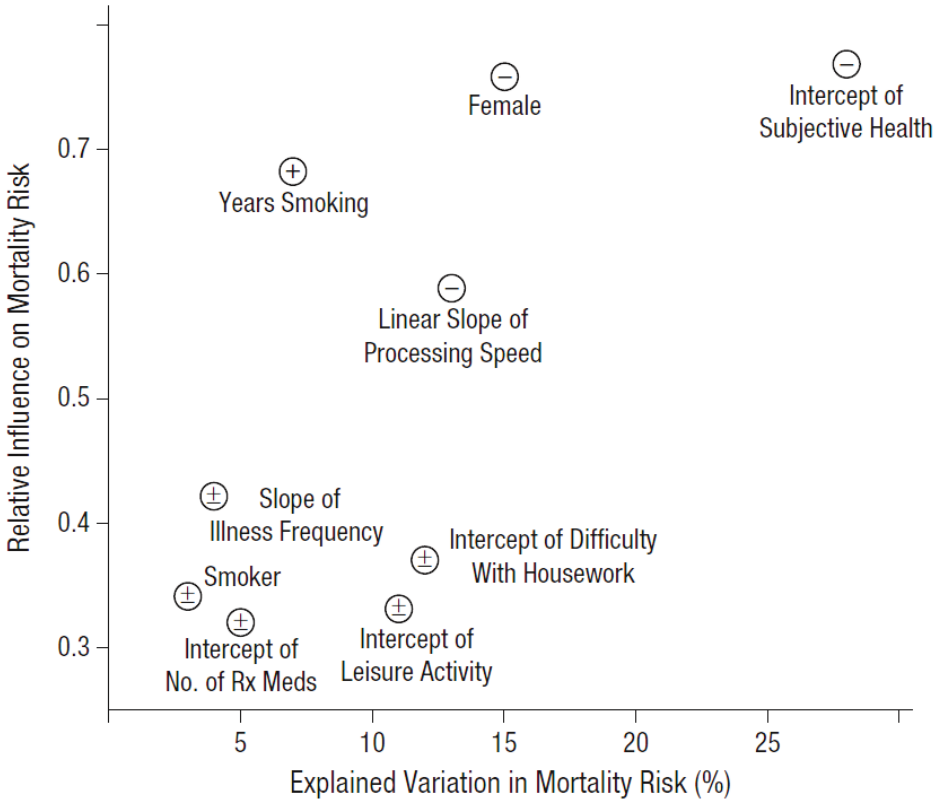
Results

	RFSA	Cox PH	
Variable	Rel. VIMP	$\Delta\chi^2$	%Δ Hazard Ratio (/*SD*)
Subjective Health (I)	.77	144	-16.2 [-28.7, -3.7]
Sex (= female)	.76	76	-33.0 [-44.4, -21.6]
Years Smoking	.68	36	11.4 [4.9, 17.9]
Processing Speed (LS)	.59	66	-10.9 [-16.8, -5.0]
CMI J: Frequency of Illness (LS)	.42	21	7.7 [-2.7, 18.1]
Difficulty Doing Housework (I)	.37	61	6.8 [-10.1, 23.7]
Smoker (= yes)	.34	15	13.9 [-3.3, 31.1]
Leisure Activity (I)	.33	58	8.6 [-4.9, 22.1]
Fluid Intelligence (LS)	.32	9	-5.9 [-10.6, -1.2]
Prescribed Medications (I)	.32	28	8.6 [-10.0, 27.2]

High VIMP, Sig. $\Delta\chi^2$ and %ΔHR

	RFSA	Cox PH	
Variable	Rel. VIMP	$\Delta\chi^2$	%Δ Hazard Ratio (/*SD*)
Subjective Health (I)	.77	144	-16.2 [-28.7, -3.7]
Sex (= female)	.76	76	-33.0 [-44.4, -21.6]
Years Smoking	.68	36	11.4 [4.9, 17.9]
Processing Speed (LS)	.59	66	-10.9 [-16.8, -5.0]
CMI J: Frequency of Illness (LS)	.42	21	7.7 [-2.7, 18.1]
Difficulty Doing Housework (I)	.37	61	6.8 [-10.1, 23.7]
Smoker (= yes)	.34	15	13.9 [-3.3, 31.1]
Leisure Activity (I)	.33	58	8.6 [-4.9, 22.1]
Fluid Intelligence (LS)	.32	9	-5.9 [-10.6, -1.2]
Prescribed Medications (I)	.32	28	8.6 [-10.0, 27.2]

Top Mortality Risk Indicators



Key Outcome

Two psychological variables, better subjective heath status and smaller life span decrements in processing speed, more strongly predicted reduced mortality risk than nearly all other observed demographic, lifestyle, and medical indices.

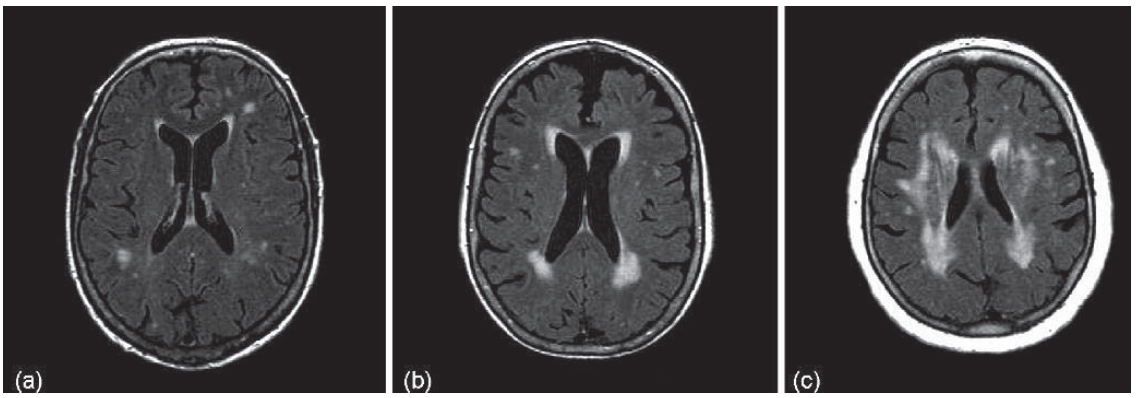
Further Interpretation

Subjective health has previously been linked to mortality risk, but not in the presence of so many other (mutually conditioned) risk factors. More specific indices (e.g., cardiovascular symptoms) may prove more informative in less healthy samples (e.g., long-term smokers).

Processing speed decrements are associated with both cardiovascular illness and cerebral white matter lesions (i.e., impaired functional connectivity). Therefore, PS may act like a biomarker of general health.

Follow-Up Study

Comparative influence of cognitive, demographic, lifestyle, & health variables on cerebral white matter hyperintensity (lesion) burden



Methodology

N = 112

Age range at time of MRI scan: 62-86 years

Longitudinal variables re-estimated within sub-sample as a function of time in study (10-15 years) rather than as a function of chronological age

Cerebral white matter lesions assessed as regional (and total) counts of hyperintensities on T1-weighted magnetic resonance images (MRI)

Random forest analysis + Generalized linear regression: WMH counts modeled as Poisson- or negative binomial-distributed

Socio-Demographic
- Age at MRI
- Gender
- Study Cohort
- Secondary education (years)
- Occupational Class
- Marital Status
- Number of Children

Tobacco and Alcohol Use
- Smoker?
- Years Smoking
- Drinker?
- Years Drinking Alcohol
- Daily Alcohol Consumption

Depression
- Geriatric Depression Scale

Daily Life Attributes
- Subjective Health
- *Prescribed Medications*
- Sleep (hr./night)
- Wake-ups (tot/night)
- Number of Hobbies
- *Difficulty Doing Housework*
- Impaired Physical Mobility
- *Leisure Activity*
- Casual Contacts (tot/wk)
- Short Conversations (tot/wk)
- Long Conversations (tot/wk)

Cognitive Abilities
- Crystallized Intelligence
- *Fluid Intelligence*
- *Processing Speed*

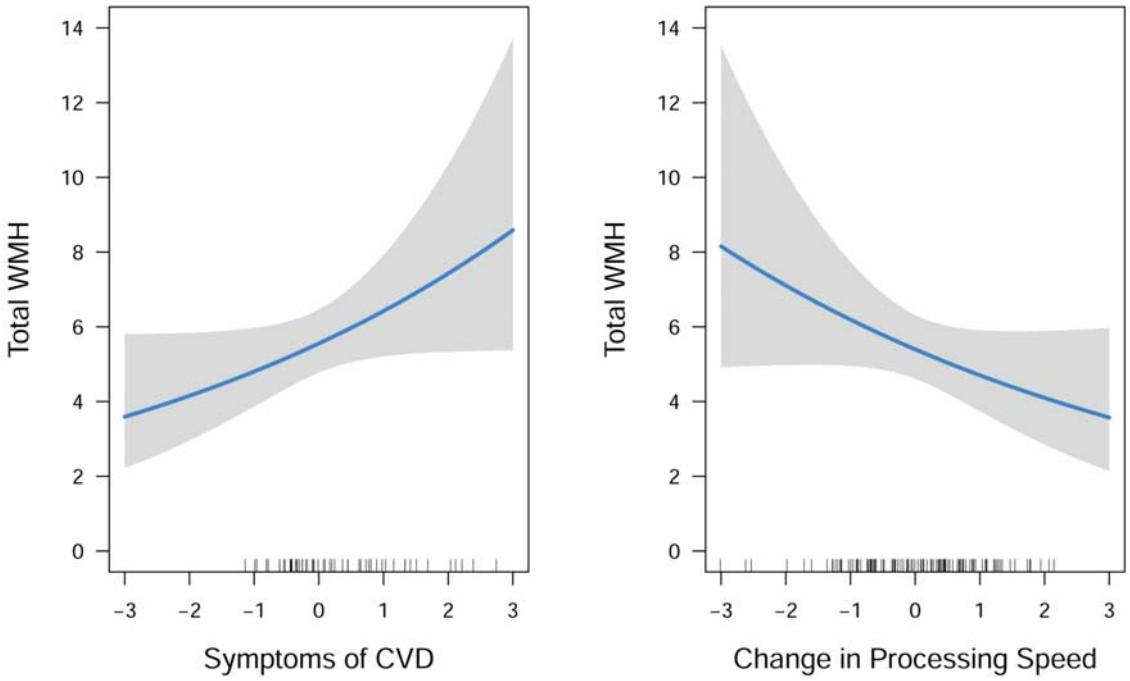
CMI Medical Symptoms (26)
- A: Eyes/Ears
- B: Nose/Throat/Respiration
- C: Cardiovascular
- D1: Teeth
- D2: Gastrointestinal/Liver
- E: Musculoskeletal
- F: Skin
- G: Nervous System
- H: Reproductive/Urinary
- I: Fatigue
- J: Frequency Illness
- K: Miscellaneous
- L: Addiction
- M: Inadequacy
- N: Depression
- O: Anxiety
- P: Sensitivity
- Q: Anger
- R: Tension
- Total symptoms (A–L , M–R)

Results (Total WMH)

	RFA	GLR (Poisson)	
Variable	Rel. VIMP	$\Delta\chi^2$	*Prevalance Rate* [95%CI]
Age in Years at MRI scan	1.00	30	1.04 [1.03, 1.06]
Processing Speed (Δ)	0.36	12	0.87 [0.80, 0.94]
Cardiovascular Symptoms (L)	0.27	18	1.17 [1.08, 1.27]
Fluid Intelligence (L)	0.10	< 1	1.00 [0.8, 1.0]

Note: For RFA, WMH were transformed via exponentiation (^2/3)

Top Predictors of Total WMH (Partial Regression Plots)



Substantive Conclusions

In general, results are consistent with the vascular hypothesis of cognitive aging, which posits that:

Vascular disease

=> Cerebral white matter degeneration

=> Functional disconnectivity (across neural networks)

=> Processing speed decrements

=> Increased mortality risk

Region specific WMH results, to be presented at a future date, paint a (slightly) more complex picture.

Methodological Implications

Different ways to aggregate data, some of which prove more theoretically informative than others

- 300 variables, factor analysis to aggregate across measures, latent growth curve models to aggregate across time/occasions
- RFA can be used to "reduce" information at the predictor-outcome level by identifying those variables with the strongest associations. Especially useful for the "small *n* large *p*" problem.
- But researchers in the social sciences will probably find it necessary to pair RFA with better known methods (e.g., multilevel and structural equation modeling, Cox proportional hazard survival analysis)

Acknowledgement

Support for this work was provided by the Swiss National Centre of Competence in Research LIVES – Overcoming vulnerability: Life course perspectives, which is financed by the Swiss National Science Foundation. We are grateful to the Swiss National Science Foundation for its financial assistance.

RFA Software Packages for R

randomForest – the original, but prone to variable selection bias

cforest – part of the "party" package in R

- Implements an unbiased algorithm for variable selection, but requires complete data (predictor and outcome variables) for optimal estimation
- Provides no built-in method for handling missing data
- Does not currently support variable importance estimation for survival models

randomForestSRC – very flexible, well-documented, optimized code that can handle large data sets. Many possibilities for working with survival models. Provides built-in mechanism for imputing missing data. Not clear if variable split-selection algorithm is biased.

Select References for Random Forest Analysis

Breiman L. (2001). Random forests, Machine Learning, 45:5-32.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., & Lauer M.S. (2008). Random survival forests. *Annals of Applied Statistics*, *2*, 841-860.

Louppe, G. (2014). *Understanding random forests: From theory to practice* (Doctoral dissertation). Retrieved from https://github.com/glouppe/phd-thesis

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods, 14*, 323-348.

Strobl, C. (2013). Data mining. In *The Oxford Handbook of Quantitative Methods* (ed. T. D. Little), pp. 678–700. New York, NY: Oxford University Press.

Example Applications from the Authors

Aichele, S., Rabbitt, P., & Ghisletta, P. (2016). Think fast, feel fine, live long: A 29-year study of cognition, health, and survival in middle-aged and older adults. *Psychological Science, 27* (4), 518–529.

Ghisletta, P., Aichele, S., & Rabbitt, P. (August, 2014). Longitudinal data mining to predict survival in a large sample of adults. In Gilli, M., González-Rodríguez, G., & Nieto-Reyes, A. (Eds.), *Proceedings of COMPSTAT 2014*: *21st International Conference on Computational Statistics* (pp. 167-175). http://www.compstat2014.org/auxil/Proceedings-COMPSTAT2014.pdf

FNSNF Swiss National Science Foundation

The National Centres of Competence in Research (NCCR)
are a research instrument
of the Swiss National Science Foundation

UNIVERSITÉ DE GENÈVE
UNIL | Université de Lausanne

A Data Mining Approach to Longitudinal Risk Assessment in Cognitive Epidemiology

Stephen Aichele & Paolo Ghisletta