UNIVERSITÀ DEGLI STUDI DI MILANO

Identifying risk factors and predicting cancer risk: past, present and future in a personal overview

Valeria Edefonti, Torino, CPO, 2024

Outline of the presentation

- **Past: tools of a dedicated epidemiologist**
  - Single exposures and cancer risk
    - Standard approaches
    - Nonlinearity in main effects and cancer risk
  - Bidimensional exposures and cancer risk: nonadditivity and nonlinearity
  - Multi-dimensional exposures and cancer risk: dietary patterns
  - Residual confounding
- **Future: machine learning in the INDACO project**
  - Nonlinearity
  - Nonadditivity
  - In the relation between (potentially complex) dietary exposures, potential confounders, and cancer risk
- **Present: machine learning for biomedical data:**
  - Comparison with statistics: two possible continua of models
  - Challanges and ethical issues

Single exposures and cancer risk: standard approaches

Table 2. Descriptive statistics on raw values of vitamin E intake (mg per day) across studies and in all the studies combined (International Head and Neck Cancer Epidemiology (INHANCE) consortium)

Study name	20%	Median	Mean	80%
Boston	5.37	7.91	9.00	11.58
Buffalo	4.47	6.90	7.78	10.45
Italy Multicenter	10.16	14.08	15.17	19.31
Japan (2001–2005)	6.08	7.42	7.77	9.26
Los Angeles	4.46	6.50	7.51	9.42
Milan (2006–2009)	8.85	11.98	12.76	16.41
MSKCC	5.05	7.22	8.84	11.34
North Carolina (2002–2006)	4.95	7.29	8.04	10.64
Switzerland	9.73	12.90	13.49	16.84
US Multicenter	3.43	4.60	4.88	6.21
All studies combined	5.37	8.30	9.73	13.48

Abbreviation: MSKCC = Memorial Sloan Kettering Cancer Center.

FULL PAPER

BJC

British Journal of Cancer (2015) 113, 182–192 | doi: 10.1038/bjc.2015.149

Keywords: head and neck cancer; INHANCE; laryngeal cancer; oral and pharyngeal cancer; vitamin E

Vitamin E intake from natural sources and head and neck cancer risk: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium

V Edefonti*[,1], M Hashibe[2], M Parpinel[3], M Ferraroni[1], F Turati[4], D Serraino[5], K Matsuo[6], A F Olshan[7], J P Zevallos[8], D M Winn[9], K Moysich[10], Z-F Zhang[11], H Morgenstern[12], F Levi[13], K Kelsey[14], M McClean[15], C Bosetti[16], S Schantz[17], G-P Yu[18], P Boffetta[19], S-C Chuang[20], Y-C A Lee[21], C La Vecchia[1] and A Decarli[1,4]

Table 3. Odds ratios (ORs)[a] of oral and pharyngeal combined and laryngeal cancers and corresponding confidence intervals (95% CIs) on vitamin E intake quintile categories (International Head and Neck Cancer Epidemiology (INHANCE) consortium)

	Oral and pharyngeal cases	Controls	OR (95% CI)[b]	$P_{studies}$[c]	Laryngeal cases	Controls	OR (95% CI)[b]	$P_{studies}$[c]
I Quintile	976	1479	1 (Reference)	0.011	315	1479	1 (Reference)	0.464
II Quintile	788	1832	0.79 (0.69–0.90)		280	1832	0.94 (0.76–1.16)	
III Quintile	704	1944	0.65 (0.56–0.74)		248	1944	0.75 (0.60–0.93)	
IV Quintile	707	1922	0.64 (0.55–0.74)		298	1922	0.93 (0.75–1.14)	
V Quintile	719	1819	0.59 (0.49–0.71)		261	1819	0.67 (0.54–0.83)	
$P_{for\ linear\ trend}$			<0.001				<0.001	

[a]Estimated from multiple logistic regression models adjusted for age, sex, education, race/ethnicity, study centre, cigarette smoking status, cigarette intensity, cigarette duration, cigar smoking status, pipe smoking status, alcohol drinking intensity and an interaction term between cigarette intensity and alcohol drinking intensity.
[b]For the oral and pharyngeal cancer, heterogeneity between studies was detected ($P<0.1$) and we reported the mixed-effects estimates derived from the corresponding generalised linear mixed model; for laryngeal cancer, there was no appreciable heterogeneity between studies and we reported the fixed-effects estimates.
[c]P for heterogeneity between studies.

Single exposures and cancer risk: nonlinearities

Table 2 Distribution of 454 endometrial cancer cases and 908 controls, and corresponding odds ratio (OR) with 95% confidence intervals (CIs)[a], according to body mass index (BMI) at diagnosis and at different ages,[b] Italy, 1992–2006

	Cases		Controls		
	No.	(%)	No.	(%)	OR (95% CI)
Height (cm)					
<160	152	(33.5)	258	(28.5)	1[c]
160–164	148	(32.6)	280	(31.0)	0.90 (0.66–1.21)
⩾165	154	(33.9)	366	(40.5)	0.71 (0.53–0.95)
χ^2 for trend (*P*-value)					5.39 (*P* = 0.02)
Weight (kg)					
<64	109	(24.0)	355	(39.1)	1[c]
64–74	145	(31.9)	311	(34.3)	1.51 (1.10–2.06)
⩾75	200	(44.1)	242	(26.7)	2.71 (1.99–3.70)
χ^2 for trend (*P*-value)					40.17 (*P*<0.01)
Body mass index (kg m^{-2})					
<20	11	(2.4)	58	(6.4)	0.56 (0.27–1.15)
20 to <25	115	(25.3)	355	(39.3)	1[c]
25 to <30	160	(35.2)	351	(38.8)	1.41 (1.05–1.90)
⩾30	168	(37.0)	140	(15.5)	4.08 (2.90–5.74)
χ^2 for trend (*P*-value)					67.95 (*P*<0.01)
BMI (kg m^{-2}) 5-Unit increase					1.89 (1.65–2.17)
Perceived body size at age 12 years					
Thinner than peers	146	(32.3)	351	(39.1)	1[c]
Same than peers	173	(38.3)	341	(38.0)	1.12 (0.85–1. 94)
Heavier than peers	133	(29.4)	206	(22.9)	1.45 (1.06–1.98)
χ^2 trend (*P*-value)					5.19 (*P* = 0.02)

Table 3 Distribution of 454 endometrial cancer cases and 908 controls, and corresponding odds ratio (OR) with 95% confidence intervals (CIs),[a] according to measures of fat distribution,[b] Italy, 1992–2006

	Cases		Controls		
	No.	(%)	No.	No.	OR (95% CI)
Waist circumference (cm)					
<84	79	(25.7)	221	(37.2)	1[c]
84–95	101	(32.9)	226	(38.1)	1.22 (0.83–1.79)
⩾96	127	(41.4)	147	(24.8)	2.68 (1. 78–4.03)
χ^2 for trend (*P*-value)					22.51 (*P*<0.01)
Hip circumference (cm)					
<100	87	(28.4)	218	(36.8)	1[c]
100 to108	96	(31.4)	204	(34.5)	1.35 (0.92–1.98)
⩾109	123	(40.2)	170	(28.7)	2.49 (1.66–3.72)
χ^2 for trend (*P*-value)					18.99 (*P*<0.01)
Waist-to-hip ratio					
<0.833	71	(23.3)	224	(37.8)	1[c]
0.833 to <0.890	129	(42.2)	177	(29.9)	2.10 (1.43–3.09)
⩾0.890	106	(34.6)	191	(32.3)	1.33 (0.89–1.97)
χ^2 for trend (*P*-value)					1.38 (*P* = 0.24)

British Journal of Cancer (2011) 104, 1207–1213

www.bjcancer.com

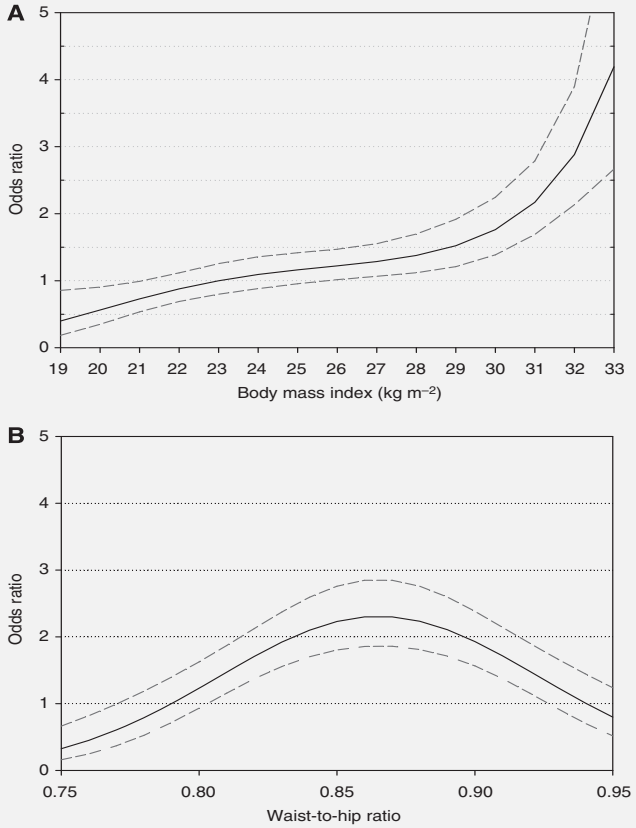


Figure 1 Estimates of odds ratios and 95% confidence intervals of endometrial cancer by body mass index at diagnosis (**A**) and waist-to-hip ratio (**B**), using cubic regression splines. Italy, 1992–2006 (Odds ratios from regression equations include terms for age, study centre, year of interview, education, smoking status, age at menarche, age at menopause, oral contraceptives use, parity, and hormone replacement therapy use. Curves are shown for best-fitting cubic spline regression models according to Akaike Information Criterion. Dashed lines represent 95% confidence intervals. Ranges represent the distribution of variables among controls from 10th to 90th percentile). Reference categories were body mass index = 23 and waist-to-hip ratio = 0.79.

Anthropometric measures at different ages and endometrial cancer risk

L Dal Maso*[,1,2], A Tavani[3], A Zucchetto[1], M Montella[4], M Ferraroni[2], E Negri[3], J Polesel[1], A Decarli[2], R Talamini[1], C La Vecchia[2,3] and S Franceschi[5]

Bidimensional exposures and cancer risk: nonadditivity

Table 3. ORs of HCC and Corresponding 95% CIs According to Family History of Liver Cancer in First-Degree Relatives*

	Cases/Controls	Model 1 OR† (95% CI)	Model 2 OR‡ (95% CI)	Model 3 OR§ (95% CI)
Number of first-degree relatives with liver cancer				
All subjects				
0	204/412	1‖	1‖	1‖
≥1	25/19	2.64 (1.39-5.02)	3.04 (1.57-5.91)	2.38 (1.01-5.58)
Males				
0	166/280	1‖	1‖	1‖
≥1	17/12	2.19 (0.99-4.81)	2.68 (1.16-6.18)	3.21 (1.13-9.10)
Females				
0	38/132	1‖	1‖	1‖
≥1	8/7	3.79 (1.25-11.46)	3.69 (1.16-11.72)	1.11 (0.21-5.78)
Type of affected relative				
No affected relatives¶	204/412	1‖	1‖	1‖
Parents	17/8	4.86 (1.99-11.87)	5.58 (2.23-14.00)	6.08 (1.99-18.62)
Siblings	9/11	1.38 (0.55-3.50)	1.60 (0.62-4.18)	0.69 (0.20-2.33)
Age of youngest affected relative#				
No affected relatives	204/412	1‖	1‖	1‖
<60	10/11	2.29 (0.93-5.69)	2.72 (1.08-6.90)	1.58 (0.46-5.40)
≥60	9/8	2.12 (0.77-5.81)	2.19 (0.79-6.11)	2.18 (0.62-7.72)
Sex of the affected relative**				
No affected relatives	204/412	1‖	1‖	1‖
Male	16/10	3.26 (1.41-7.54)	3.28 (1.39-7.71)	2.29 (0.80-6.58)
Female	9/9	1.41 (0.86-2.39)	1.67 (1.00-2.78)	1.59 (0.80-3.14)
Family history of liver cancer using FHscore††				
Minimal-risk	204/412	1‖	1‖	1‖
Low-/intermediate-risk	12/10	1.83 (0.75-4.47)	1.89 (0.76-4.72)	1.42 (0.43-4.72)
High-risk	13/9	3.82 (1.56-9.36)	4.91 (1.95-12.33)	3.87 (1.20-12.55)
P value for trend		≤0.01	≤0.01	0.02

*Italy, 1999-2002.
†Estimated from unconditional multiple logistic regression models adjusted for age, sex, and center.
‡Further adjusted for education, alcohol drinking, and smoking habits.
§Further adjusted for HBsAg and/or anti-HCV positivity.
‖Reference category.
¶One subject reported both a parent and a sibling affected by liver cancer.
#The sum does not add to the total because of some missing values on age at liver cancer diagnosis in first-degree relatives.
**One subject had the mother and a brother affected by liver cancer
††There were two missing values for the FHscore.

Family History of Liver Cancer and Hepatocellular Carcinoma

Federica Turati,[1,2] Valeria Edefonti,[2] Renato Talamini,[3] Monica Ferraroni,[2] Matteo Malvezzi,[1,2,4] Francesca Bravi,[1,2] Silvia Franceschi,[5] Maurizio Montella,[6] Jerry Polesel,[3] Antonella Zucchetto,[3] Carlo La Vecchia,[1,2,7] Eva Negri,[1] and Adriano Decarli[2,4]

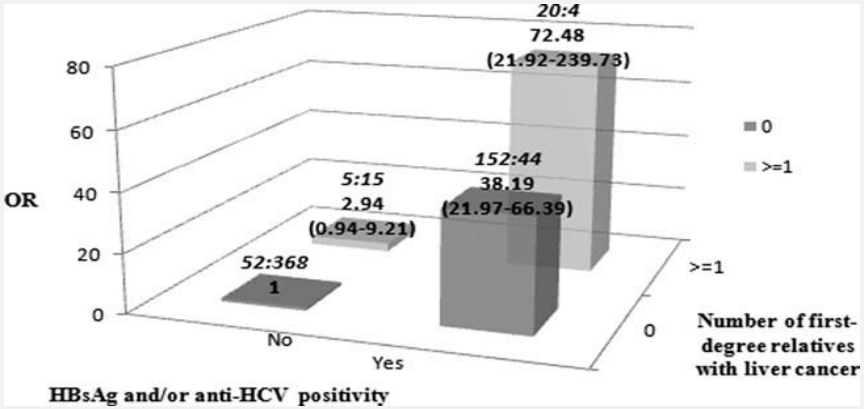


Fig. 1. Number of cases and controls, ORs* and 95% CIs of hepatocellular carcinoma according to chronic hepatitis and family history of liver cancer, measured by the standard method. Italy, 1999-2002. *Adjusted for age, sex, center, education, alcohol drinking, and smoking habits.

UNIVERSITÀ DEGLI STUDI DI MILANO

Bidimensional nonlinear exposures and cancer risk

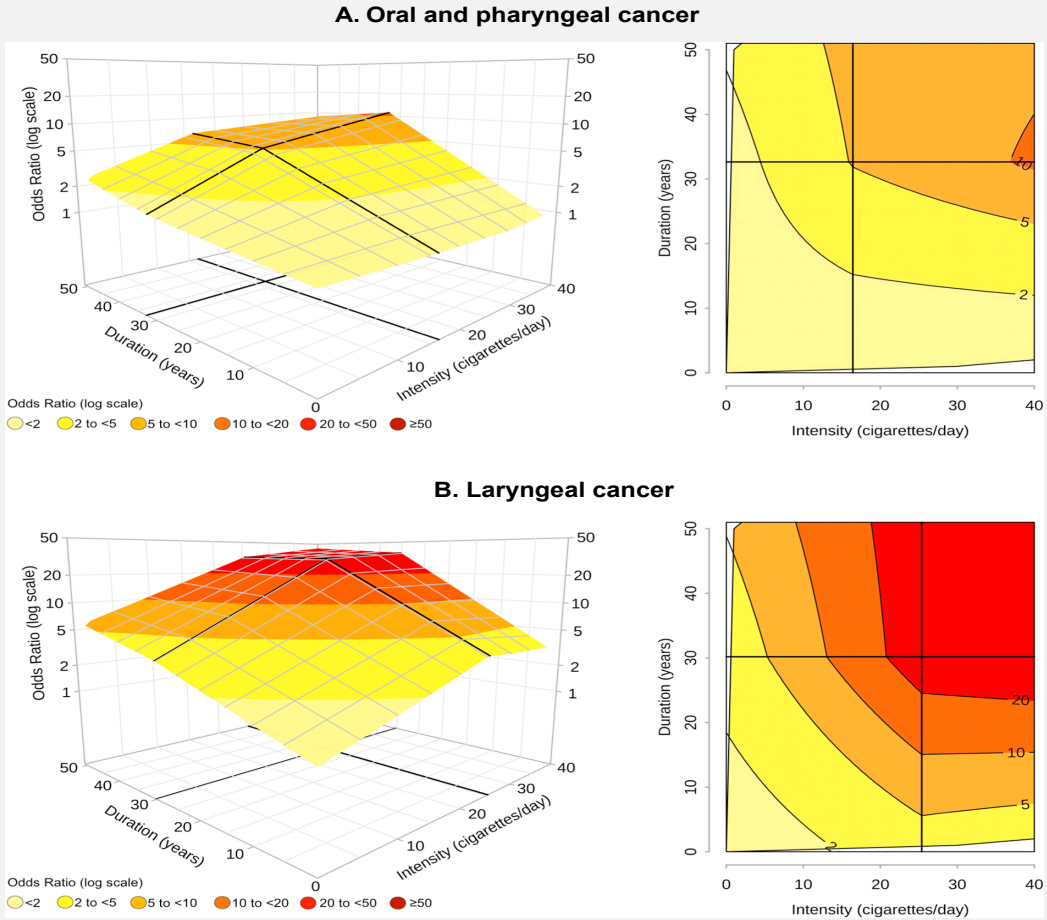


Fig. 2. Odds ratios[a,b] of oral and pharyngeal cancer and laryngeal cancer in current smokers, for the joint effect of intensity (cigarettes/day) and duration (years) of cigarette smoking estimated through bivariate spline models. INHANCE consortium. [a]Fitted models included adjustment for age, sex, race, study, education, drinking status, drinking intensity, and drinking duration. The reference category was defined as "Never smokers". [b]On the grid, black thicker lines represent knot locations: 16 cigarettes/day and 33 years of duration for oral and pharyngeal cancer and 25 cigarettes/day and 30 years of duration for laryngeal cancer, respectively. Dark grey lines in contour plots (right) indicate iso-risk curves at defined risk levels.

Contents lists available at ScienceDirect

Oral Oncology

journal homepage: www.elsevier.com/locate/oraloncology

Joint effects of intensity and duration of cigarette smoking on the risk of head and neck cancer: A bivariate spline model approach

Gioia Di Credico[a,b], Valeria Edefonti[c,*], Jerry Polesel[d], Francesco Pauli[b], Nicola Torelli[b], Diego Serraino[d], Eva Negri[c], Daniele Luce[f], Isabelle Stucker[g], Keitaro Matsuo[h], Paul Brennan[i], Marta Vilensky[j], Leticia Fernandez[k], Maria Paula Curado[l], Ana Menezes[m], Alexander W. Daudt[n], Rosalina Koifman[o], Victor Wunsch-Filho[p], Ivana Holcatova[q], Wolfgang Ahrens[r,s], Pagona Lagiou[t], Lorenzo Simonato[u], Lorenzo Richiardi[v], Claire Healy[w], Kristina Kjaerheim[x], David I. Conway[y], Tatiana V. Macfarlane[z], Peter Thomson[aa], Antonio Agudo[ab], Ariana Znaor[i], Leonardo F. Boaventura Rios[ac], Tatiana N. Toporcov[p], Silvia Franceschi[d], Rolando Herrero[i], Joshua Muscat[ad], Andrew F. Olshan[ae], Jose P. Zevallos[af], Carlo La Vecchia[c], Deborah M. Winn[ag], Erich M. Sturgis[ah], Guojun Li[ah], Eleonora Fabianova[ai], Jolanda Lissowska[aj], Dana Mates[ak], Peter Rudnai[al], Oxana Shangina[am], Beata Swiatkowska[an], Kirsten Moysich[ao], Zuo-Feng Zhang[ap], Hal Morgenstern[aq], Fabio Levi[ar], Elaine Smith[as], Philip Lazarus[at], Cristina Bosetti[au], Werner Garavello[av], Karl Kelsey[aw], Michael McClean[ax], Heribert Ramroth[ay], Chu Chen[az], Stephen M. Schwartz[az], Thomas L. Vaughan[az], Tongzhang Zheng[ba], Gwenn Menvielle[bb], Stefania Boccia[bc,bd], Gabriella Cadoni[be,bf], Richard B. Hayes[bg], Mark Purdue[ag], Maura Gillison[bh], Stimson Schantz[bi], Guo-Pei Yu[bj], Hermann Brenner[bk,bl,bm], Gypsyamber D'Souza[bn], Neil D. Gross[bo], Shu-Chun Chuang[bp], Paolo Boffetta[bq], Mia Hashibe[br], Yuan-Chin Amy Lee[bs], Luigino Dal Maso[d]

A B S T R A C T

Objectives: This study aimed at re-evaluating the strength and shape of the dose-response relationship between the combined (or joint) effect of intensity and duration of cigarette smoking and the risk of head and neck cancer (HNC). We explored this issue considering bivariate spline models, where smoking intensity and duration were treated as interacting continuous exposures.

Materials and Methods: We pooled individual-level data from 33 case-control studies (18,260 HNC cases and 29,844 controls) participating in the International Head and Neck Cancer Epidemiology (INHANCE) consortium. In bivariate regression spline models, exposures to cigarette smoking intensity and duration (compared with never smokers) were modeled as a linear piecewise function within a logistic regression also including potential confounders. We jointly estimated the optimal knot locations and regression parameters within the Bayesian framework.

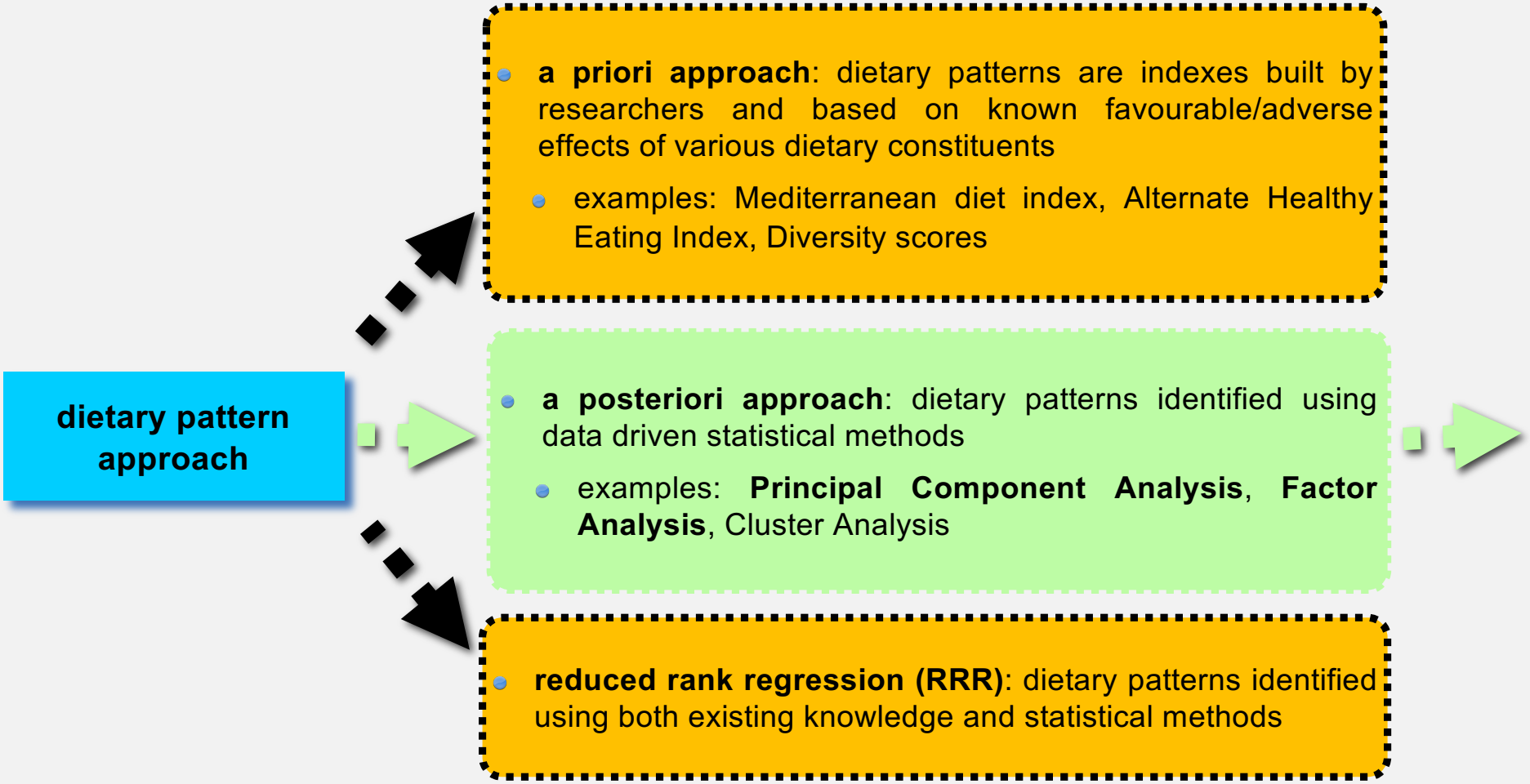
Results: For oral-cavity/pharyngeal (OCP) cancers, an odds ratio (OR) > 5 was reached after 30 years in current smokers of ~20 or more cigarettes/day. Patterns of OCP cancer risk in current smokers differed across strata of alcohol intensity. For laryngeal cancer, ORs > 20 were found for current smokers of ≥20 cigarettes/day for ≥30 years. In former smokers who quit ≥10 years ago, the ORs were approximately halved for OCP cancers, and ~1/3 for laryngeal cancer, as compared to the same levels of intensity and duration in current smokers.

Conclusion: Referring to bivariate spline models, this study better quantified the joint effect of intensity and duration of cigarette smoking on HNC risk, further stressing the need of smoking cessation policies.

Multi-dimensional exposures and cancer risk: dietary patterns

- Free-living individuals eat meals consisting of a **variety of foods with complex combinations of interacting nutrients**
- **Dietary patterns** are one or **more combined variables summarizing multiple interacting dietary components** and thus capturing the **cumulative exposure to different dietary components**
- **Dietary patterns may have stronger effects on health/disease risk than any single component**

Multi-dimensional exposures and cancer risk: dietary patterns/2



Multi-dimensional exposures and cancer risk: dietary patterns/3

- Compared to one-dimensional *a priori* DPs, ***a posteriori*** **DPs describe actual dietary behavior**
- **A continuous score is assigned to each subject representing his/her cumulative exposure on one or more** profiles that we called **patterns**
- **The statistical analysis of the associetion between dietary patterns and cancer risk works like if they were single nutrients**

Multi-dimensional exposures and cancer risk: dietary patterns/4

Table 2. Factor loading matrix and explained variances for the four major dietary patterns identified by factor analysis

Nutrient	Animal products	Vitamins and fiber	VUFA	Starch-rich
Animal protein	**0.80**	0.10	0.41	0.23
Vegetable protein	0.15	0.39	0.29	**0.80**
Cholesterol	**0.72**	0.07	0.41	0.30
Saturated fatty acids	0.56	0.15	0.50	0.41
Monounsaturated fatty acids	0.20	0.29	**0.72**	0.28
Linoleic acid	0.19	0.16	**0.71**	0.33
Linolenic acid	0.33	0.27	**0.68**	0.34
Other polyunsaturated fatty acids	0.48	−0.02	**0.75**	−0.04
Soluble carbohydrates	0.40	**0.66**	0.02	0.17
Starch	0.18	0.11	0.26	**0.88**
Sodium	0.41	0.06	0.16	**0.80**
Calcium	**0.65**	0.34	0.03	0.28
Potassium	0.42	**0.76**	0.29	0.28
Phosphorus	**0.70**	0.37	0.31	0.45
Iron	0.42	0.48	0.39	0.37
Zinc	**0.63**	0.29	0.45	0.47
Thiamin (vitamin B1)	0.53	0.51	0.30	0.45
Riboflavin (vitamin B2)	**0.76**	0.47	0.10	0.26
Vitamin B6	0.53	0.58	0.41	0.29
Total folate	0.40	**0.71**	0.22	0.28
Niacin	0.54	0.37	0.47	0.21
Vitamin C	0.12	**0.85**	0.13	−0.11
Retinol	0.47	0.08	0.03	0.00
β-Carotene equivalents	0.04	**0.67**	0.20	0.02
Lycopene	−0.05	0.26	0.49	0.32
Vitamin D	0.54	0.04	0.54	−0.23
Vitamin E	0.08	0.53	**0.74**	0.22
Total fiber (Englyst)	0.06	**0.85**	0.15	0.31
Proportion of explained variances (%)	21.67	20.30	18.02	15.10
Cumulative explained variances (%)	21.67	41.97	59.99	75.09

NOTE: Estimates from a PCFA done on 28 nutrients. Loadings ≥0.63 are shown in boldface.

Nutrient Dietary Patterns and Gastric Cancer Risk in Italy

Paola Bertuccio,[1,2] Valeria Edefonti,[2] Francesca Bravi,[1,2] Monica Ferraroni,[3] Claudio Pelucchi,[1] Eva Negri,[1] Adriano Decarli,[2,4] and Carlo La Vecchia[1,2]

Table 3. OR of gastric cancer and corresponding 95% CIs on quartiles of factor scores from a PCFA

Dietary pattern	Quartile category, OR (95% CI)				P_{trend}*
	I†	II	III	IV	
Animal products	1	1.08 (0.64-1.80)	1.47 (0.90-2.40)	2.13 (1.34-3.40)	0.0003
Vitamins and fiber	1	0.84 (0.53-1.32)	1.00 (0.64-1.56)	0.60 (0.37-0.99)	0.0861
VUFA	1	0.84 (0.53-1.34)	0.89 (0.56-1.42)	0.89 (0.56-1.42)	0.7325
Starch-rich	1	1.37 (0.83-2.25)	1.37 (0.82-2.28)	1.67 (1.01-2.77)	0.0463

NOTE: Estimates from a logistic regression model conditioned on age and sex and adjusted for quinquennia of period of interview, education, body mass index, tobacco smoking, and family history of gastric cancer. Results refer to the composite model including all the four factors simultaneously.
*P value for linear trend.
†Reference category.

Multi-dimensional exposures and cancer risk: dietary patterns/5

- There is emerging evidence that **DPs and disease may have nonlinear relations**
- **Improper specification of models due to erroneous/incomplete exposure characterization or assumptions can lead to masked or spurious associations and biased estimates**

Residual confounding

- **Dense correlations between dietary components and with confounding factors can make it difficult** to ascertain the most relevant dietary exposures and **to address residual confounding**
- Even when confounders are appropriately specified in models, **residual confounding can remain if unspecified nonadditivity/nonlinearity is present**

Machine learning: the promise

- «Machine learning approaches problems as a doctor progressing through residency might: by learning rules from data. Starting with patient-level observations, algorithms sift through vast numbers of variables, looking for combinations that reliably predict outcomes
- In one sense, this process is similar to that of traditional regression models: there is an outcome, covariates, and a statistical function linking the two, **but where machine learning shines is in handling enormous numbers of predictors and combining them in nonlinear and highly interactive ways**» (N Engl J Med 2016; 375: 1216–1219)

Machine learning: the promise/2

- **Machine learning** may be used in nutritional epidemiology to explore:

  - **more complex**
  - **numerous dietary variables in models**

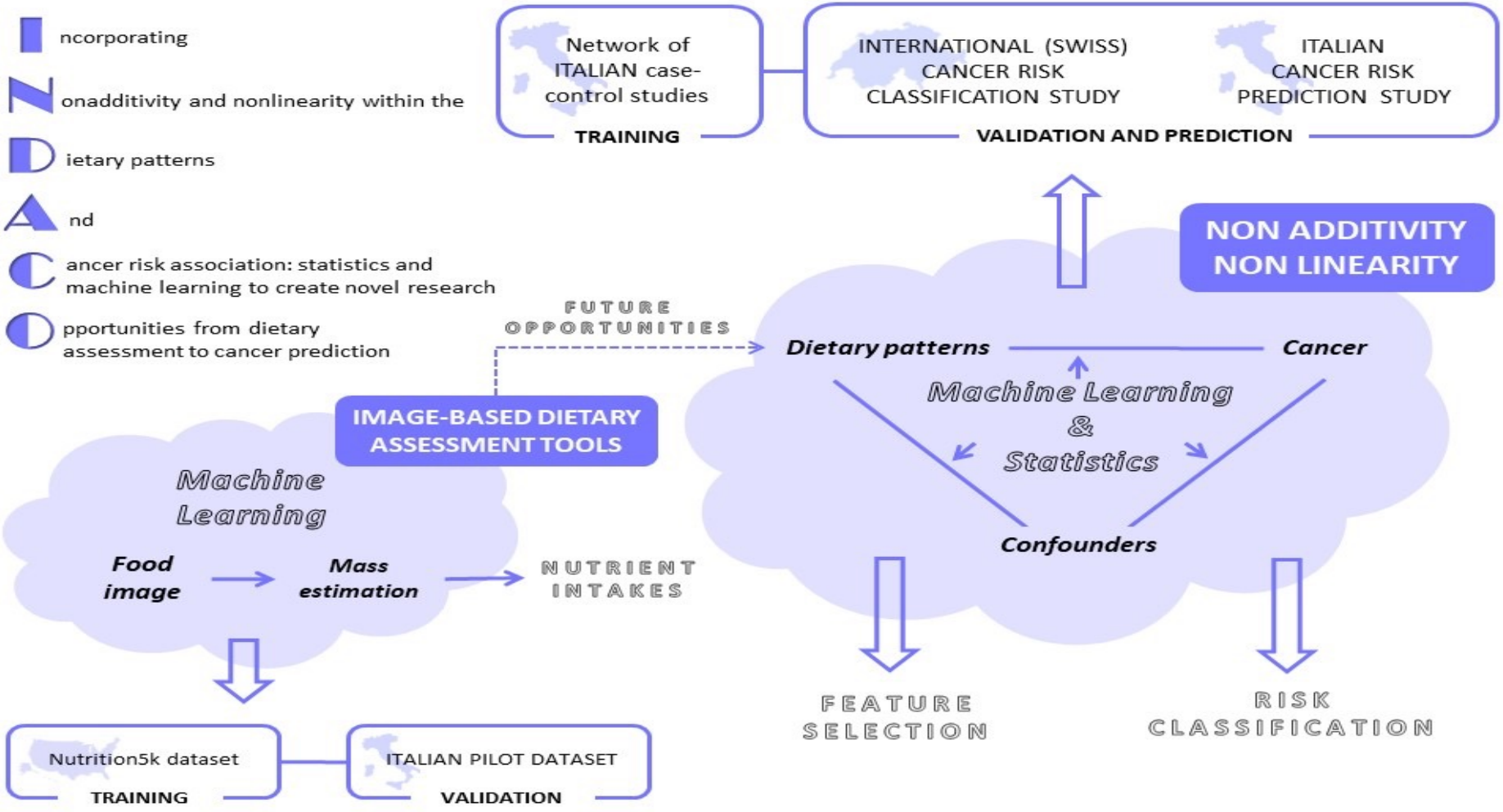
  - and/or

  - **nonlinear**
  - **nonadditive relations**
  - **between diet, other confounders, and cancer risk**

PRIN 2022 – INDACO: objectives

- Explore nonlinearity and nonadditivity of dietary patterns -> cancer risk relation by developing novel machine learning (ML) and statistical approaches
- Evaluate the identified cancer risk prediction/classification models with
  - previously collected database of Swiss case-control studies on diet and cancer
  - newly collected database of middle-aged, healthy university employees from Milan and Udine
- Explore ML approaches for image-based dietary assessment
- Transfer results from ML US-based image classification to a newly collected pilot study of Italian food images and recipes

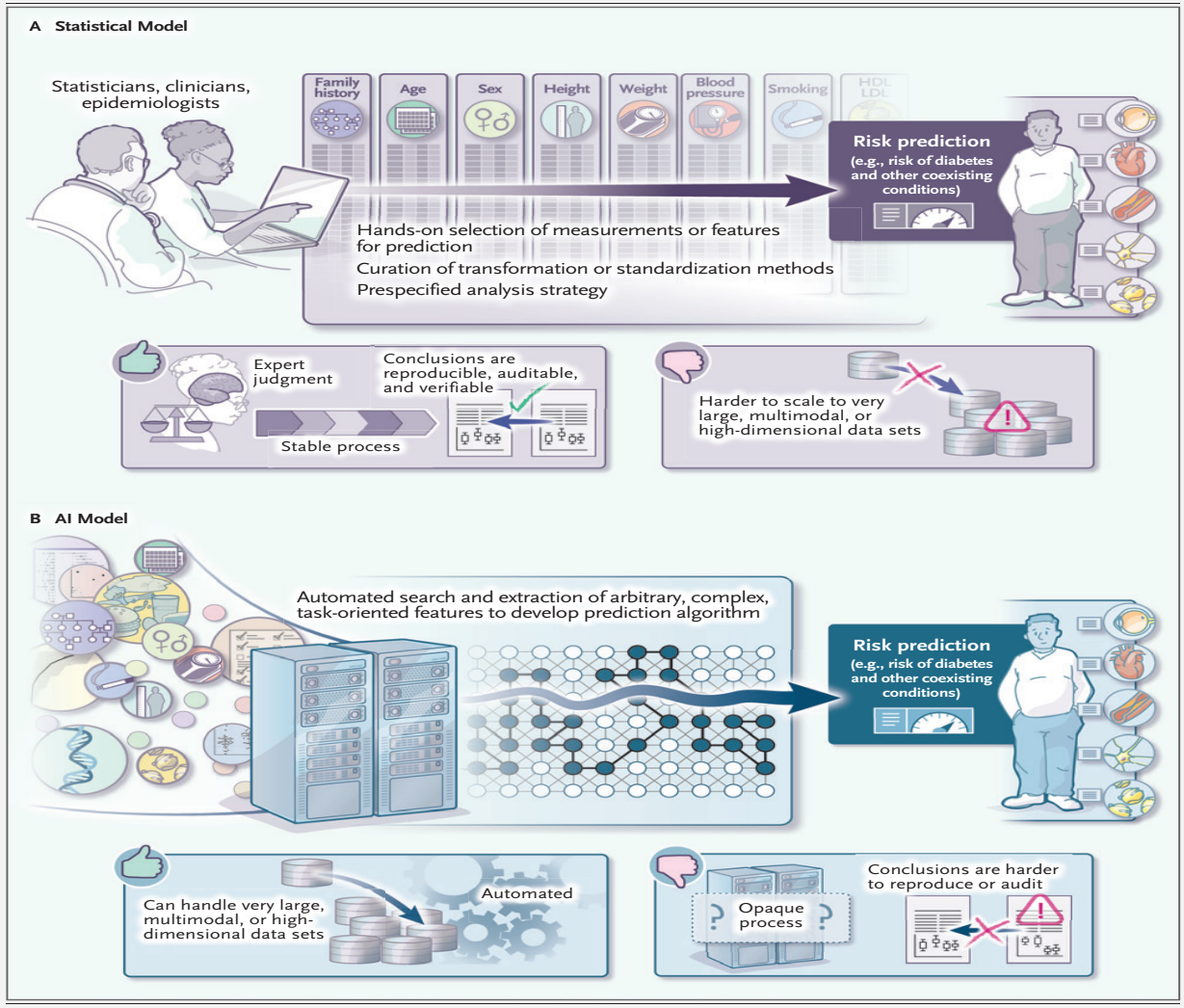
PRIN 2022 – INDACO: graphical overview



Machine learning for biomedical data

- Machine learning is a subfield of artificial intelligence where profiles are derived from data with little human input
- This contrasts with statistical techniques that rely more on human knowledge and emphasize a theoretical approach to uncertainty

Statistics and machine learning



The NEW ENGLAND JOURNAL of MEDICINE

REVIEW ARTICLE

AI IN MEDICINE
Jeffrey M. Drazen, M.D., *Editor*, Isaac S. Kohane, M.D., Ph.D., *Guest Editor*, and Tze-Yun Leong, Ph.D., *Guest Editor*

Where Medical Statistics Meets Artificial Intelligence

David J. Hunter, M.B., B.S., and Christopher Holmes, Ph.D.

Statistics and machine learning/2

Table 1. Similarities and Differences between Artificial Intelligence and Conventional Statistics.

Feature	Artificial Intelligence Methods	Conventional Statistical Methods
Prior hypotheses	Agnostic or very general	Specific; often categorized as primary, secondary, and exploratory
Techniques (examples)	Random forests, neural networks, XGBoost	Parametric and nonparametric comparisons between groups; regression and survival models with linear predictors
Stability (end-to-end)	Analyses are more prone to instability and variability as a result of application domains (e.g., multimodal data integration) and user choices in algorithm specification (e.g., architecture in deep learning)	Stable analyses that follow prespecification of a statistical analysis plan with minimal available user-defined choices in model specification
Applications	Analysis of images, outputs from monitors, massive data sets (e.g., electronic health records, natural language processing)	Data with a smaller number of predictors, tabular data, randomized trials
Purpose	Pattern discovery; automatic feature representation; feature reduction to a smaller, more manageable set; prediction models	Statistical inference and testing of specific factors for departure from a null hypothesis, control of confounding and ascertainment bias, quantification of uncertainty
Reproducibility	Often internal (i.e., performed with original data set); cross-validation or split samples	Ideally external (i.e., performed with "new" data); formal tests of significance against null hypotheses
Barriers	Increasingly, use of proprietary algorithms not available to other researchers; lack of clarity in reporting	Slow progress in sharing of primary data to allow others to check or extend results
Interpretability	Often black-box; automatic algorithmic feature engineering introduces opaqueness	Explicit features, clear number of free parameters and degrees of freedom
Equity	Data-driven feature learning susceptible to biases present in data, compounding health inequities	Less flexible, more explicit (interpretable) models, which are more easily checked for equity if relevant data are available

Machine learning: feature representation learning

- **Feature representation learning** - the most impressive and distinguishing aspect of machine learning - is its **automated ability to search and extract arbitrary, complex, task-oriented features from data**
- Features are algorithmically engineered from data during a **training phase** in order to uncover data transformations that are correct for the learning task

Machine learning: feature representation learning/2

- **Optimality** is measured by means of an "**objective function**" quantifying how well the AI model is performing the task at hand
- **AI models can search through potentially billions of nonlinear covariate transformations to reduce a large number of variables to a smaller set of task-adapted features**
- **AI algorithms largely remove the need for analysts to prespecify features for prediction or manually curate transformations of variables**

Machine learning: feature representation learning/3

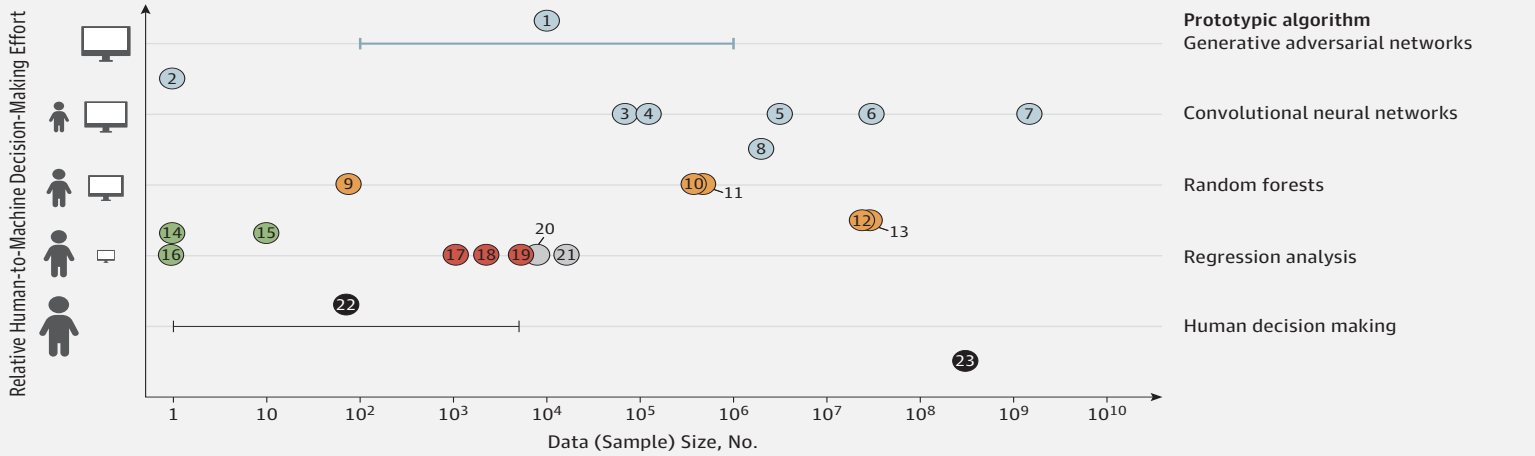
- **The trained AI models can engineer data-adaptive features that are beyond the scope of features that humans can engineer**
- **Such features can be hard to interpret and lack common sense in the use of background knowledge and qualitative checks that statisticians bring to bear in deciding on a feature set to use in a model**
- AI models are often **unable to trace the evidence line from data to features**, making auditability and verification challenging

Machine learning spectrum: a continuum of models

- We can **imagine an algorithm as existing along a continuum between fully human-guided vs fully machine-guided data analysis**
- To understand the degree to which an algorithm can said to be of machine learning requires **understanding how much of its assumptions** (structure or parameters) **were predetermined by humans**
- The trade-off between human specification of properties vs learning properties from data is the **machine learning spectrum**

Machine learning spectrum: a continuum of models/2

Figure. The Axes of Machine Learning and Big Data



Deep learning

1. Generative adversarial networks (2014)
2. Google AlphaGo Zero (2017)
3. ATM check readers (1998)
4. Google diabetic retinopathy (2016)
5. ImageNet computer vision models (2012-2017)
6. Google AlphaGo (2015)
7. Facebook Photo Tagger (2015)
8. Prediction of 1-y all-cause mortality (2017)

Classic machine learning

9. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling (2002)
10. EHR-based CV risk prediction (2017)
11. Netflix Prize winner (2006)
12. Google Search (1998)
13. Amazon product recommendation (2003)

Expert AI systems

14. MYCIN (1975)
15. CASNET (1982)
16. DXplain (1986)

Risk calculators

17. CHA_2DS_2-VASc Score for atrial fibrillation stroke risk (2017)
18. MELD end-stage liver disease risk score (2001)
19. Framingham CV risk score (1998)

Randomized Clinical Trials

20. Celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis (2002)
21. Use of estrogen plus progestin in healthy postmenopausal women (2002)

Other

22. Clinical wisdom
23. Mortality rate estimates from US Census (2010)

Traditional clinical studies analyze data from hundreds or thousands of patients using a carefully designed statistical model and thus are low on the machine learning spectrum. Deep learning models are at the top of the spectrum. At the very top are generative adversarial networks, which can learn to generate new images by examining a large database of existing images. See the Supplement for details including supporting references and expansions of abbreviations.

Machine learning spectrum: a continuum of models/3

- **When human effort was used to define properties, it would place low on the machine learning spectrum** (#19)
- **High on the machine learning spectrum are deep learning models,** stunningly complex networks of artificial neurons designed to create accurate models directly from raw data (#4)

Machine learning spectrum: a continuum of models/5

- The **flexibility** offered by the high end of the spectrum requires **vast amounts of computational resources** must be used to develop and deploy these algorithms
- While algorithms high on the spectrum are often very flexible, **they are often uninterpretable and function mostly as "black boxes"**
- In contrast, **algorithms lower on the spectrum often produce outputs that are easier for humans** to understand and interpret

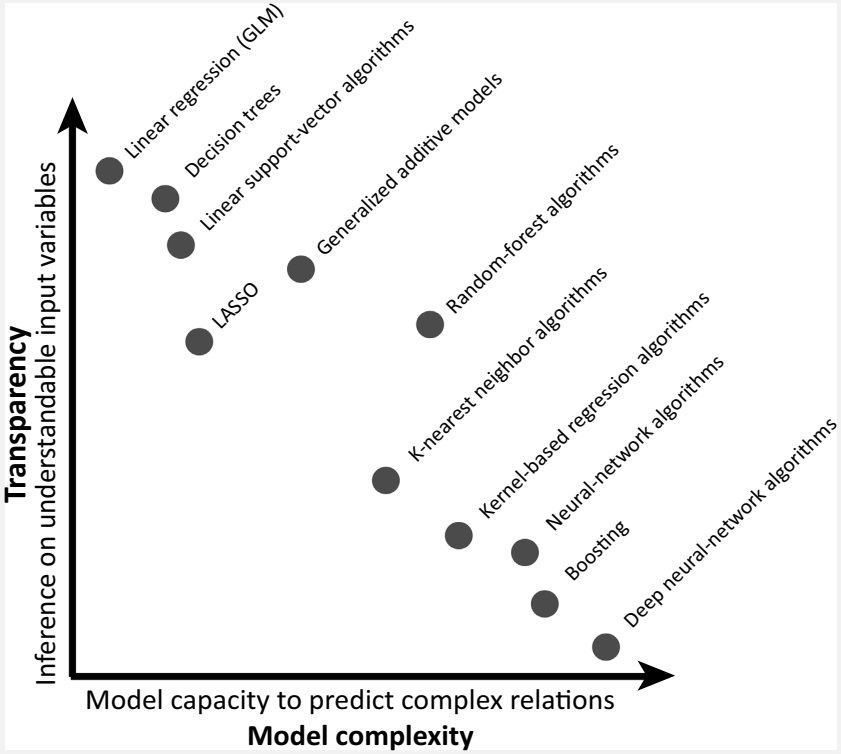
Inference–Prediction: a second continuum of models

- Statistics and machine learning models can be posed on a second continuum, based on the **main motivation of the analysis, which can be doing inference or prediction at the extremes**
- The **inferential regime** prioritizes statements about the **relevance of each individual input variable;** the **predictive regime** prioritizes the **relevance of the output of the model for precise forecasting**
- **Machine learning is especially well suited to, and largely designed for, large-scale prediction tasks**

Inference–Prediction: a continuum of models/2

- **Inferring** new scientific insight is often about answering: **which input variable within a given dataset is an important contributor to the outcome?** The investigator is interested in understanding **the way in which an outcome is affected by a change in the input variables**
- **Predictive modeling** describes **what 'does' happen**, but does **not equally well address the question of 'how' and may be less apt for the question of 'why'**

Inference–Prediction: a continuum of models/3



Trends in Neurosciences

Figure 1. The Trade-Off between Model Transparency, Which Allows Scientific Understanding, and Theoretical Model Capacity, Which Affords Sophisticated Predictions. Neuroscience and biomedicine have had a long-dominating focus on scientific insight by using simple and thus transparent models. Such approaches are well suited to work towards the goal of inference regarding mechanistic understanding. This goal is epistemologically distinct from, and sometimes practically incompatible with, maximizing predictive power. The pragmatic goal of optimizing predictive accuracy can exploit large datasets even at the cost of opting for black box models that cannot easily be interrogated. In practice, the actual ratio between transparency and predictability depends on the specific analytical tool being used and the particular dataset at hand. Abbreviations: GLM, generalized linear models; LASSO, least absolute shrinkage and selection operator: a recently introduced constrained regression for high-dimensional data analysis, which is a special instance of GLM.

Trends in Neurosciences

CellPress REVIEWS

Opinion

Exploration, Inference, and Prediction in Neuroscience and Biomedicine

Danilo Bzdok[1,2,3,*,@] and John P.A. Ioannidis[4,5,*]

Challenges of machine learning

- **Quantity of input data:**
  - Machine learning algorithms are highly "**data hungry**," often requiring millions of observations to reach acceptable performance levels
  - It is often difficult to know the **optimal sample size** for a particular prediction-oriented clinical research program beforehand; **reasons include the unknown complexity of the aspired prediction function, the amount of relevant input variables, and noise in the data**

Challenges of machine learning/2

- **Quality of input data:**
  - Input data should be **unambiguously defined and measured**
  - In **noisy data**, advanced pattern-learning algorithms struggle to identify reproducible signatures among the measured variables: **the more complex the predictive model, the higher its susceptibility to random variation**
  - **Biases in data collection can substantially affect both performance and generalizability**; **private companies** spend resources to amass high-quality, unbiased data to feed their algorithms, and **existing data in electronic health records or claims databases need careful curation and processing** before they are usable

Challenges of machine learning/3

- **Performance evaluation:**
  - **Choosing a measure that is appropriate for the context** (e.g..area under the ROC curve, specificity, sensitivity) **is vitally important**, since accuracy in one of these measures may not translate to accuracy in another and may not **relate to a clinically meaningful measure of performance or safety**
  - **Prediction performance needs to be better than what can be achieved using existing clinical methods** for diagnosis and monitoring

Challenges of machine learning/4

- **Overfitting and unstable estimates:**
  - Algorithms might "**overfit" predictions to spurious correlations in data**
  - **Multicollinear, correlated predictors** could produce **unstable estimates**
  - Either possibility can lead to **overly optimistic estimates of model accuracy** and exaggerated claims about real-world performance
- **Reproducibility and internal validation: overinterpretation?**
  - **The use of regularization and controlled stochastic optimization of model parameters during training can help prevent overfitting** but also means that **algorithms have poorly defined notions of statistical degrees of freedom and the number of free parameters**
  - **Cross-validation and held-out samples** are provided to mimic true out-of-sample performance, with the trade-off that **the amount of data available for discovery is reduced**

Challenges of machine learning/5

- **Generalizability and independent validation: overinterpretation?**
  - **Overfitting and unstable estimates must be addressed by testing models on truly independent validation data sets**, from different populations or periods that played no role in model development
  - **Problems in the model-fitting stage, whatever their cause, will show up as poor performance in the validation stage**
  - **Generalizability to different groups of individuals and different ethnicities** that did not contribute to model building is **important** per se

Challenges of machine learning/6

- **Causality in observational studies**:
  - The usual common-sense caveats about **confusing correlation with causation apply**
  - They become **even more important as researchers begin including millions of variables in statistical models**
- **Successful predictive models, clinical outcomes and ethics:**
  - **Predictive successes can result in better patient management and clinical outcomes as far as effective interventions are available** (e.g., Alzheimer's disease)
  - **The potential for false positive results is increased under machine learning approaches** unless rigorous procedures to assess the reproducibility of findings are incorporated
  - **New reporting guidelines and recommendations for artificial intelligence in medical science have been established to ensure greater trust and generalizability of conclusions**

Challenges of machine learning/7

- **Optimal reporting of information: large scale clinical trials**

Introduction				
Background and rationale	6a	Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention	SPIRIT-AI 6a (i) Extension	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare professionals, patients, public).
			SPIRIT-AI 6a (ii) Extension	Describe any pre-existing evidence for the AI intervention.
	6b	Explanation for choice of comparators		
Objectives	7	Specific objectives or hypotheses		
Trial design	8	Description of trial design including type of trial (eg, parallel group, crossover, factorial, single group), allocation ratio, and framework (eg, superiority, equivalence, non-inferiority, exploratory)		
Methods: Participants, interventions, and outcomes				
Study setting	9	Description of study settings (eg, community clinic, academic hospital) and list of countries where data will be collected. Reference to where list of study sites can be obtained	SPIRIT-AI 9 Extension	Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting.
Eligibility criteria	10	Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centres and individuals who will perform the interventions (eg, surgeons, psychotherapists)	SPIRIT-AI 10 (i) Elaboration	State the inclusion and exclusion criteria at the level of participants.
			SPIRIT-AI 10 (ii) Extension	State the inclusion and exclusion criteria at the level of the input data.
Interventions	11a	Interventions for each group with sufficient detail to allow replication, including how and when they will be administered	SPIRIT-AI 11a (i) Extension	State which version of the AI algorithm will be used.
			SPIRIT-AI 11a (ii) Extension	Specify the procedure for acquiring and selecting the input data for the AI intervention.
			SPIRIT-AI 11a (iii) Extension	Specify the procedure for assessing and handling poor quality or unavailable input data.
			SPIRIT-AI 11a (iv) Extension	Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users.
			SPIRIT-AI 11a (v) Extension	Specify the output of the AI intervention.
			SPIRIT-AI 11a (vi) Extension	Explain the procedure for how the AI intervention's output will contribute to decision-making or other elements of clinical practice.

RESEARCH METHODS AND REPORTING

Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension

Samantha Cruz Rivera,[1,2] Xiaoxuan Liu,[2,3,4,5,6] An-Wen Chan,[7] Alastair K Denniston,[1,2,3,4,5,8] Melanie J Calvert,[1,2,6,9,10,11] On behalf of the SPIRIT-AI and CONSORT-AI Working Group

BMJ: first published as 10.

Challenges of machine learning/8

- **Optimal reporting of information**

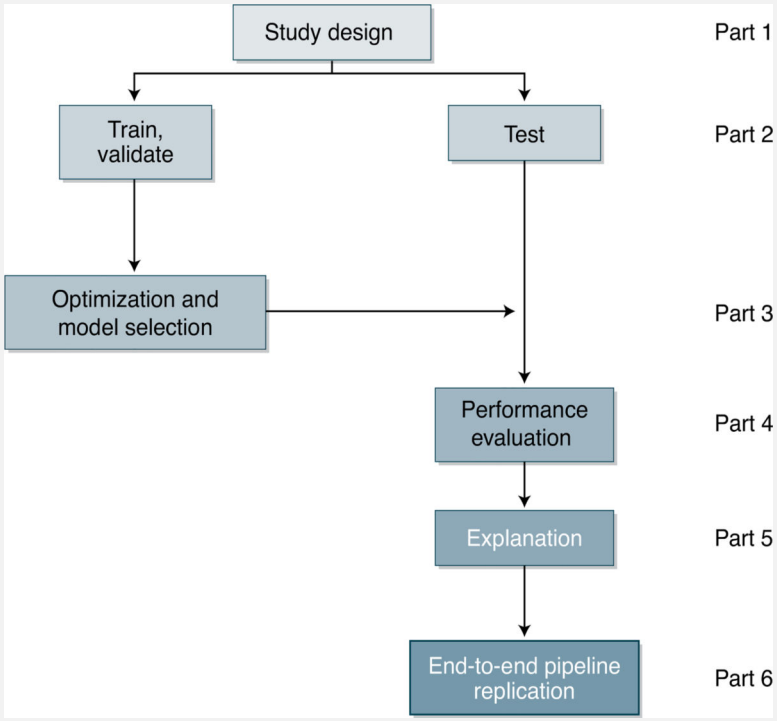


Fig. 1 |.
A schematic representation of the six components of a clinical AI study.

Published in final edited form as:
Nat Med. 2020 September ; 26(9): 1320–1324. doi:10.1038/s41591-020-1041-y.

Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist

Beau Norgeot[1], **Giorgio Quer**[2], **Brett K. Beaulieu-Jones**[3], **Ali Torkamani**[2], **Raquel Dias**[2], **Milena Gianfrancesco**[4], **Rima Arnaout**[1], **Isaac S. Kohane**[3], **Suchi Saria**[5,6], **Eric Topol**[2], **Ziad Obermeyer**[7], **Bin Yu**[8], **Atul J. Butte**[1,✉]

General conclusions/1

- Much of **the skills of a trained statistician/epidemiologist involve factors that cannot be captured by data-driven artificial intelligence algorithms**
- **Bringing these skills within a "human-in-the-loop" development** (in which artificial intelligence supports and assists expert human judgment) **will highlight gaps** to be addressed
- Human experts should be useful in **carefully specifying objective functions** for training and evaluation and **exploring the consequences of the applications of machine learning**

General conclusions/2

- As more control is ceded to algorithms, it is important to note that **these new algorithmic decision-making tools come with no guarantees of fairness, equitability, or even veracity**
- Even with the best machine learning algorithms the maxim of "**garbage in, garbage out**" remains true
- Whether an algorithm is high/low on the machine learning spectrum, best analytic practices must be used to **ensure that the end result is robust and valid**

General conclusions/3

- **The checking of artificial intelligence-supported findings is particularly important in** the emerging field of **generative artificial intelligence through self-supervised learning**, such as large language models and medical science chatbots that may be used for medical note taking in electronic health records
- Researchers should find that **delicate balance** between wishing to **learn** as much as possible **from data** while ensuring that data-driven **conclusions are accurate, robust, and reproducible**

General conclusions/4

- Although **intellectual property rights for commercial artificial intelligence products** may exist, practices that medical scientists should pay careful attention to in planning machine learning studies include **releasing all code and providing clear statements on model fitting and held-out data used for reporting of accuracy so as to facilitate external assessment of the reproducibility of findings**