

# Privacy Preserving Machine Learning

SAV Après-Midi, 31 August 2023

Daniel Meier, Swiss Re

Juan Ramón Troncoso-Pastoriza, Tune Insight

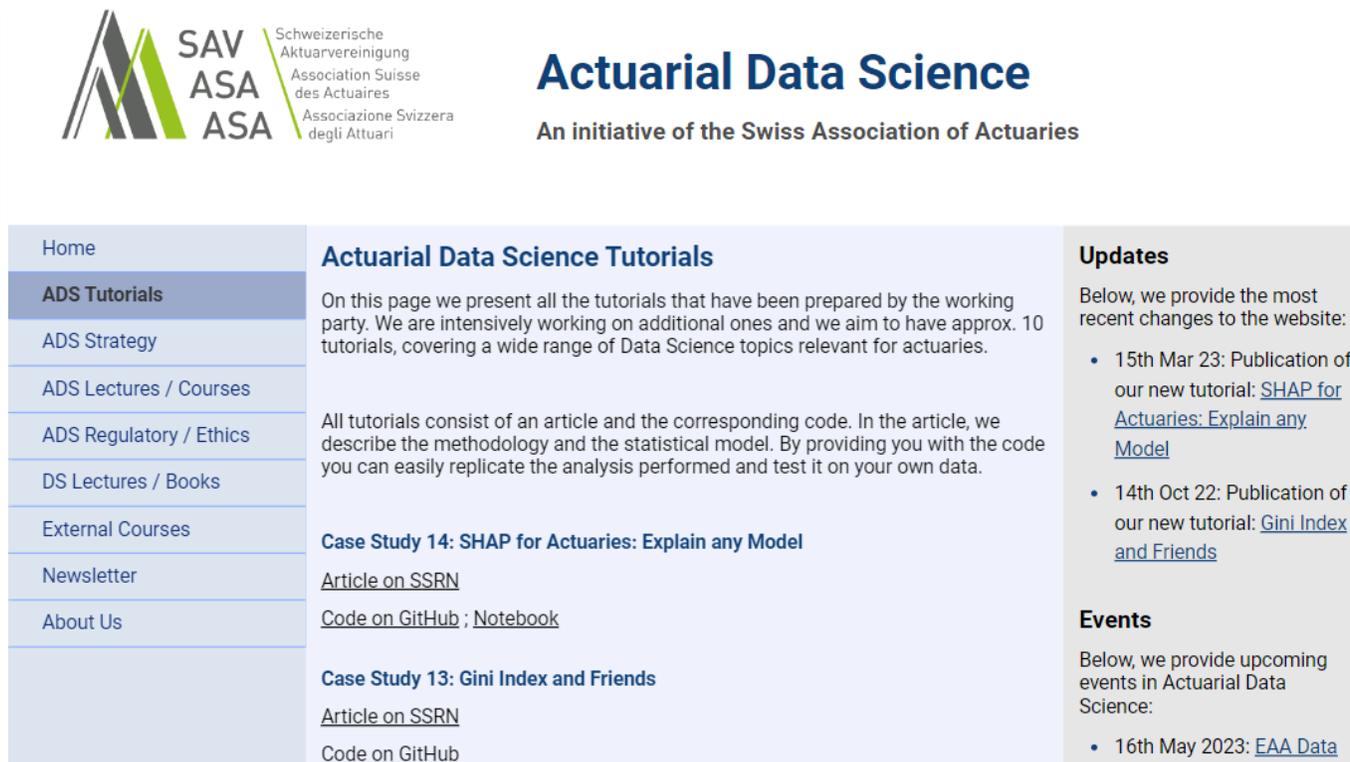
## Key take-aways

1. Limited data availability for sensitive personal (Life & Health) data in practice (e.g., nFADP, 1 September 2023)
2. For sufficiently large and dense datasets, ML/DL methods outperform traditional models, creating value for policyholders and insurance companies
3. Privacy preserving methods can help to access data

# Agenda

- Creation of synthetic health datasets
- Introducing 3 models to create health risk scores: Logistic regression, Cox regression, neural networks
- Homomorphic encryption

Paper and code soon available at [actuarialdatascience.org](https://actuarialdatascience.org)



**Actuarial Data Science**  
An initiative of the Swiss Association of Actuaries

**Home**  
**ADS Tutorials**  
ADS Strategy  
ADS Lectures / Courses  
ADS Regulatory / Ethics  
DS Lectures / Books  
External Courses  
Newsletter  
About Us

**Actuarial Data Science Tutorials**  
On this page we present all the tutorials that have been prepared by the working party. We are intensively working on additional ones and we aim to have approx. 10 tutorials, covering a wide range of Data Science topics relevant for actuaries.

All tutorials consist of an article and the corresponding code. In the article, we describe the methodology and the statistical model. By providing you with the code you can easily replicate the analysis performed and test it on your own data.

**Case Study 14: SHAP for Actuaries: Explain any Model**  
[Article on SSRN](#)  
[Code on GitHub](#) ; [Notebook](#)

**Case Study 13: Gini Index and Friends**  
[Article on SSRN](#)  
[Code on GitHub](#)

**Updates**  
Below, we provide the most recent changes to the website:

- 15th Mar 23: Publication of our new tutorial: [SHAP for Actuaries: Explain any Model](#)
- 14th Oct 22: Publication of our new tutorial: [Gini Index and Friends](#)

**Events**  
Below, we provide upcoming events in Actuarial Data Science:

- 16th May 2023: [EAA Data](#)

## (Publicly) available health datasets

- CPRD, <https://cprd.com/data>
- MIMIC, <https://physionet.org/about/database/>
- IPUMS, <https://healthsurveys.ipums.org/>
- NHANES, <https://www.cdc.gov/nchs/nhanes/>
- Nightingale, <https://docs.nightingalescience.org/>
- UK Biobank, <https://www.ukbiobank.ac.uk/>
- IHME, <https://ghdx.healthdata.org/>
- ...
- See also [longitudinal study](#) for other health datasets

- Often, access is restricted to academic institutions and/or limited to a pre-defined research topic
- Data volumes (and density) rather too low for ML
- Access to more data sources (e.g., hospitals, GPs, insurance companies, etc.) – in a privacy preserving manner – is needed

# Health risk scores, e.g., QRISK3 providing 10-year risk of a cardio-vascular disease (CVD)

**ClinRisk**  **Welcome to the QRISK<sup>®</sup>3-2018 risk calculator <https://qrisk.org>**

This calculator is only valid if you do not already have a diagnosis of coronary heart disease (including angina or heart attack) or stroke/transient ischaemic attack.

**About you**

Age (25-84):

Sex:  Male  Female

Ethnicity:

UK postcode: leave blank if unknown

Postcode:

**Clinical information**

Smoking status:

Diabetes status:

Angina or heart attack in a 1st degree relative < 60?

Chronic kidney disease (stage 3, 4 or 5)?

Atrial fibrillation?

On blood pressure treatment?

Do you have migraines?

Rheumatoid arthritis?

Systemic lupus erythematosus (SLE)?

Severe mental illness? (this includes schizophrenia, bipolar disorder and moderate/severe depression)

On atypical antipsychotic medication?

Are you on regular steroid tablets?

A diagnosis of or treatment for erectile dysfunction?

Leave blank if unknown

Cholesterol/HDL ratio:

Systolic blood pressure (mmHg):

Standard deviation of at least two most recent systolic blood pressure readings (mmHg):

**Body mass index**

Height (cm):

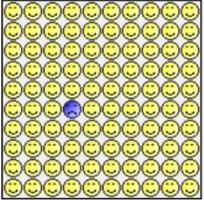
Weight (kg):

**Your results**

Your risk of having a heart attack or stroke within the next 10 years is:

**0.6%**

In other words, in a crowd of 100 people with the same risk factors as you, 1 are likely to have a heart attack or stroke within the next 10 years.



**Risk of  
a heart attack or stroke**

Your score has been calculated using estimated data, as some information was left blank.

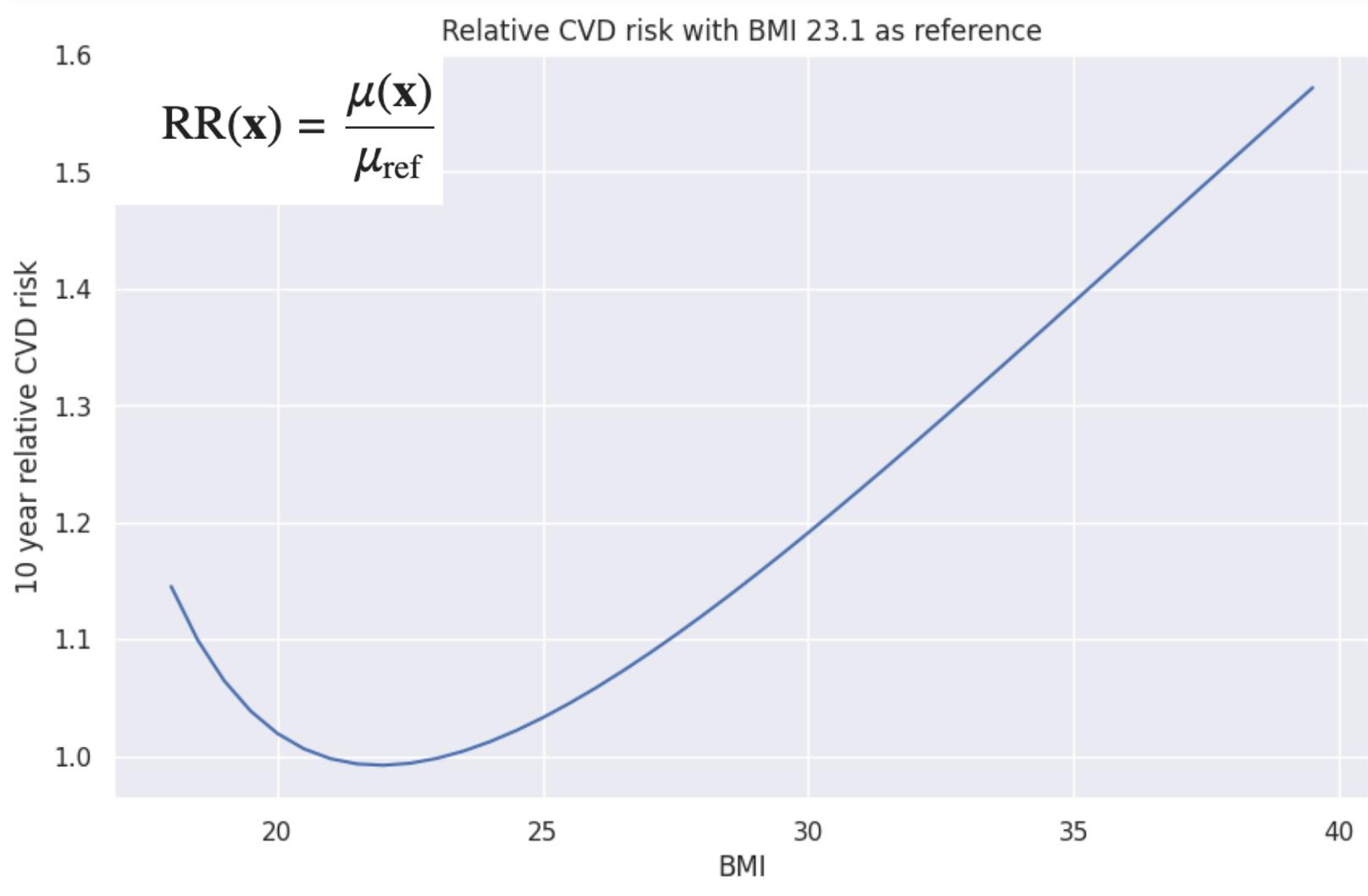
Your body mass index was calculated as 23.15 kg/m<sup>2</sup>.

**How does your 10-year score compare?**

Your score	
Your 10-year QRISK <sup>®</sup> 3 score	0.6%
The score of a healthy person with the same age, sex, and ethnicity*	0.7%
Relative risk**	0.9
Your QRISK <sup>®</sup> 3 Healthy Heart Age***	35

\* This is the score of a healthy person of your age, sex and ethnic group, i.e. with no adverse clinical indicators and a cholesterol ratio of 4.0, a stable systolic blood pressure of 125, and BMI of 25.  
\*\* Your relative risk is your risk divided by the healthy person's risk.  
\*\*\* Your QRISK<sup>®</sup>3 Healthy Heart Age is the age at which a healthy person of your sex and ethnicity has your 10-year QRISK<sup>®</sup>3 score.

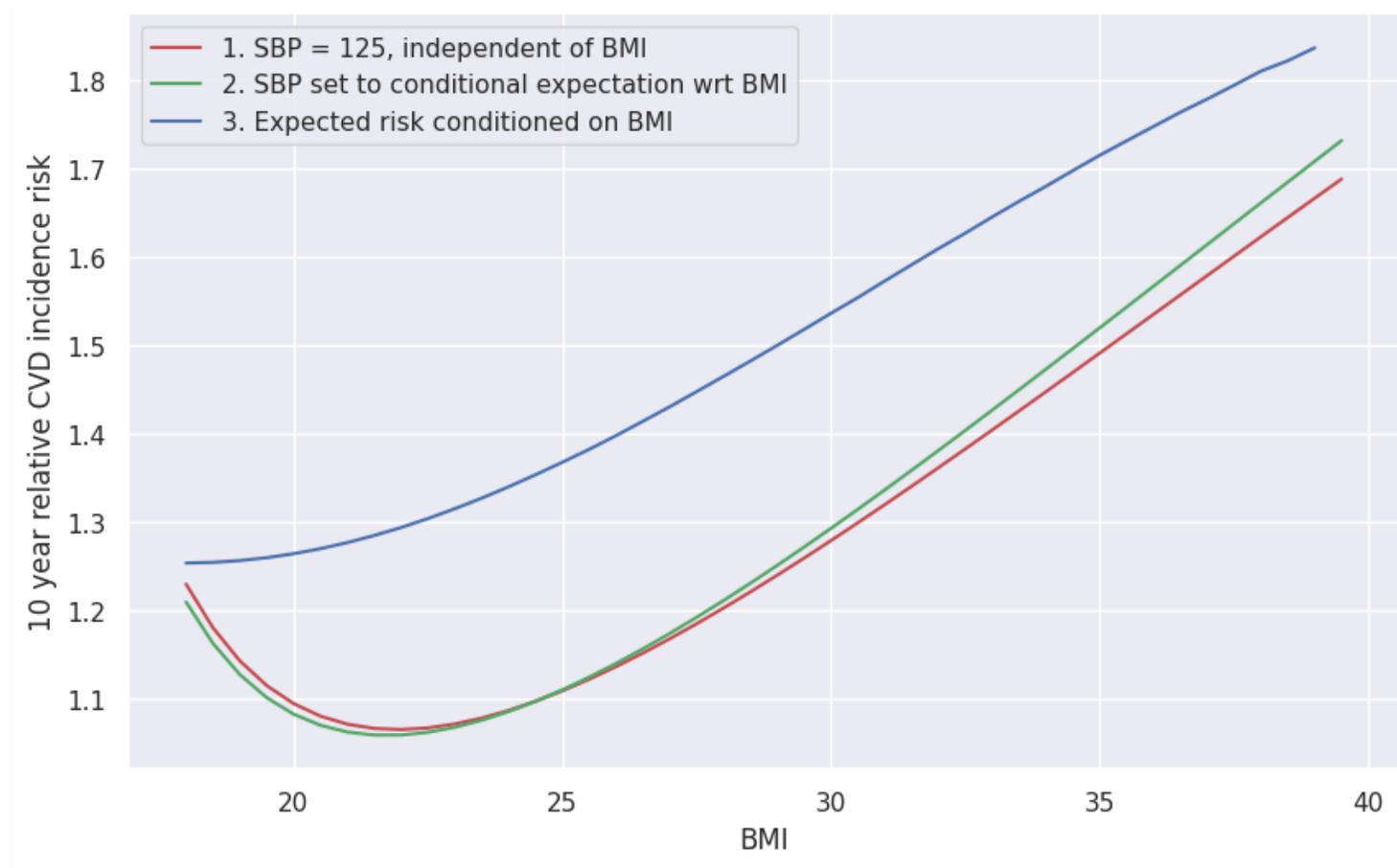
## Relative risk with respect to a reference person of same age, gender



Country	Mean BMI females	Mean BMI males
Samoa	33.5	29.9
USA	28.8	28.8
UK	27.1	27.5
Germany	25.6	27.0
Italy	25.2	26.8
France	24.6	26.1
Switzerland	23.8	26.7
Japan	21.7	23.6

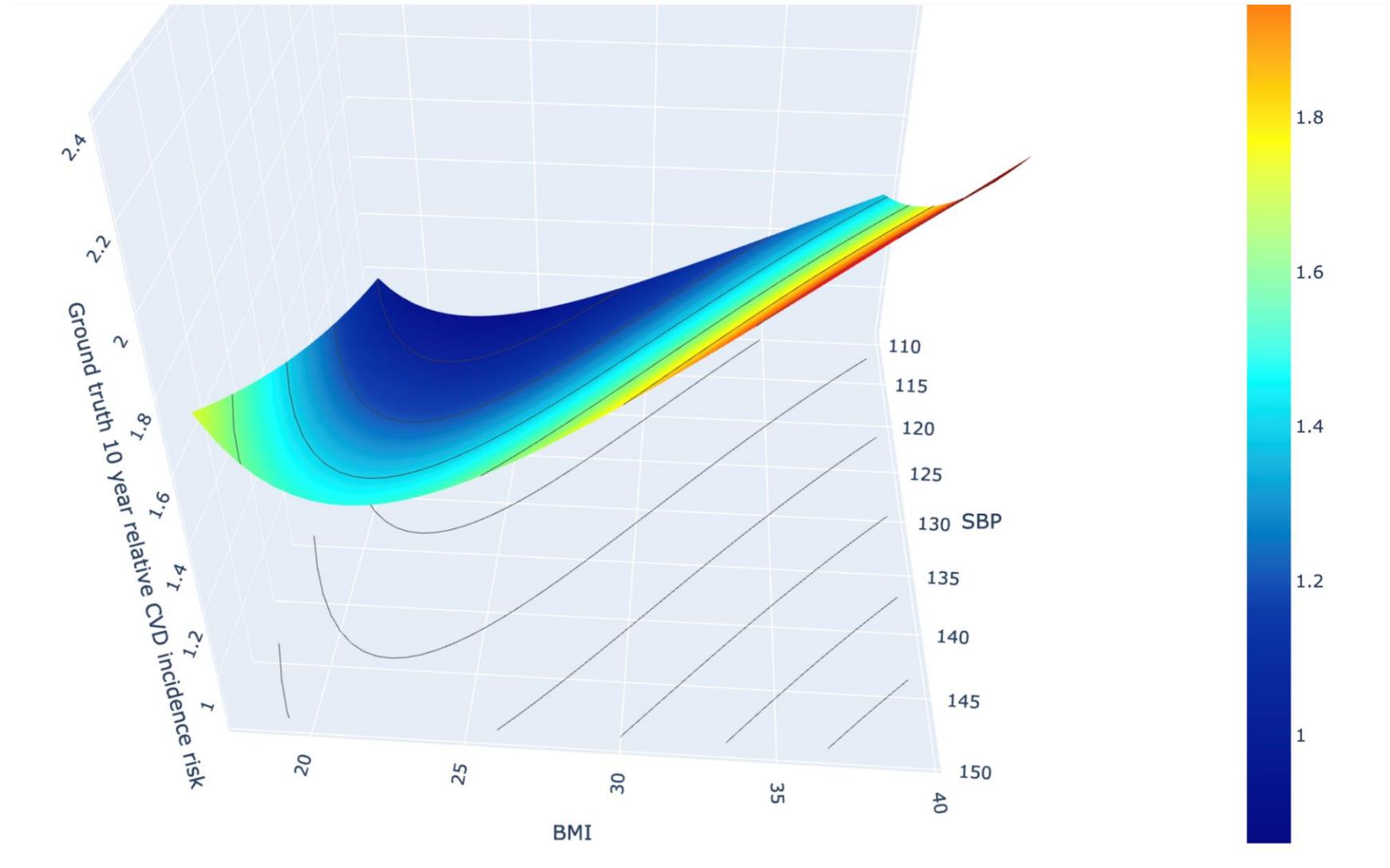
Country	Life exp. females	Life exp. males
Samoa	75.5	71.3
USA	81.5	76.5
UK	83.3	79.6
Germany	83.5	78.8
Italy	85.4	81.1
France	85.6	79.8
Switzerland	85.6	81.9
Japan	87.4	81.4

## Various risk factors like BMI, systolic blood pressure (SBP) impact relative risk



1. What is the risk of a person with a given BMI and all other attributes equal to the reference person,  $\mu(\text{BMI}, \text{SBP}_{\text{ref}})$ ?
2. What is the risk of a person with a given BMI, and SBP set to the conditional expectation of SBP given BMI,  $\mu(\text{BMI}, \mathbb{E}[\text{SBP}|\text{BMI}])$ ?
3. What is the expected risk of a person with a given BMI,  $\mathbb{E}[\mu(\mathbf{x})|\text{BMI}]$ ?
4. What is the (causally implied) risk of the reference person when changing BMI,  $\mathbb{E}[\mu(\mathbf{x})|do(\text{BMI})]$ ?

# Various risk factors like BMI, systolic blood pressure (SBP) impact relative risk



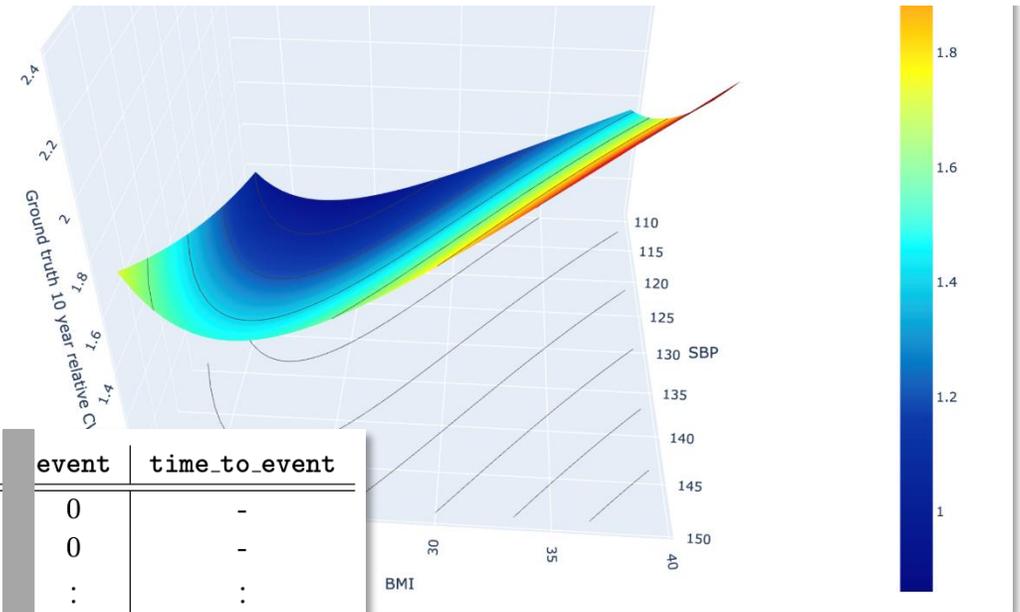
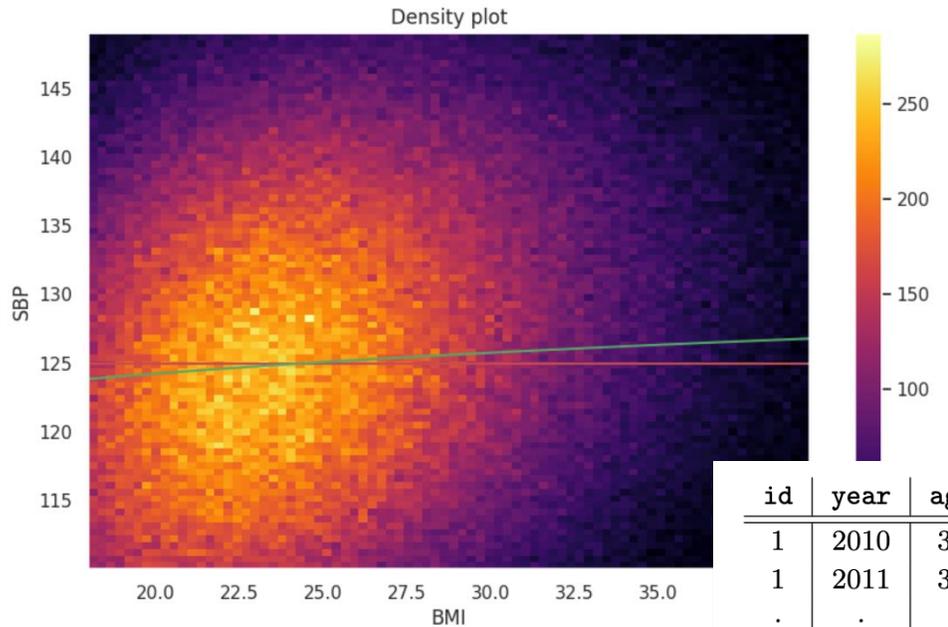
## Creation of a synthetic health dataset

- **id**: an ID to uniquely identify a person,
- **year**: observation year of health information,
- **age**: age of the person at time **year**,
- **gender**: male (0)/female (1),
- **bmi**: body-mass-index (BMI), unit  $\text{kg}/\text{m}^2$ ,
- **sbp**: systolic blood pressure (SBP), unit mmHg,
- **sd\_sbp**: standard deviation of systolic blood pressure measurements, unit mmHg,
- **tcl\_hdl\_ratio**: total cholesterol level (TCL) divided by high-density lipoprotein level (HDL),
- **num1**, **num2**, **num3**: 3 generic numeric health risk factors without specifying their meaning explicitly, e.g., stepcounts, triglycerides, resting heartrate, etc.
- **binary**: a generic binary health risk factor, e.g., smokers, foreign born, etc.,

# Creation of a synthetic health dataset

$$\begin{pmatrix} \text{sbp} \\ \log(\text{bmi}) \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 125 \\ 3.2 \end{pmatrix}, \begin{pmatrix} 15^2 & 15 \cdot 0.25\rho \\ 15 \cdot 0.25\rho & 0.25^2 \end{pmatrix} \right)$$

$$\begin{aligned} \mu^*(\mathbf{x}) &= \exp \left( \log(\text{QRISK3}(\mathbf{x})) + r(\text{age}) \right) \\ &+ 16(\text{num1} - 0.5)^4 + 4(\text{num2} - 0.5)^2 \text{num3} + \text{num3} + \text{binary} - 1.65 \end{aligned}$$



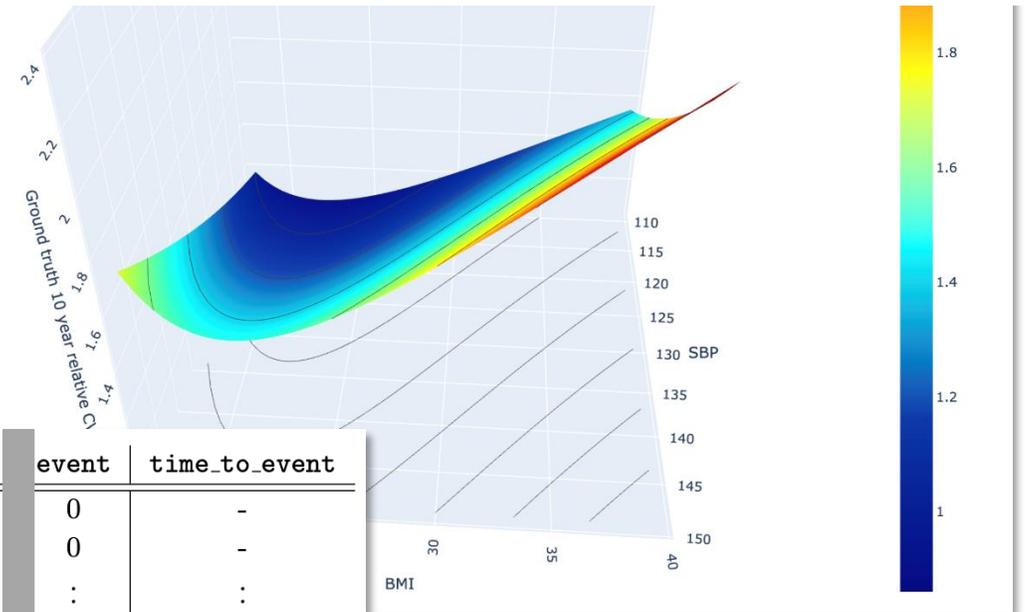
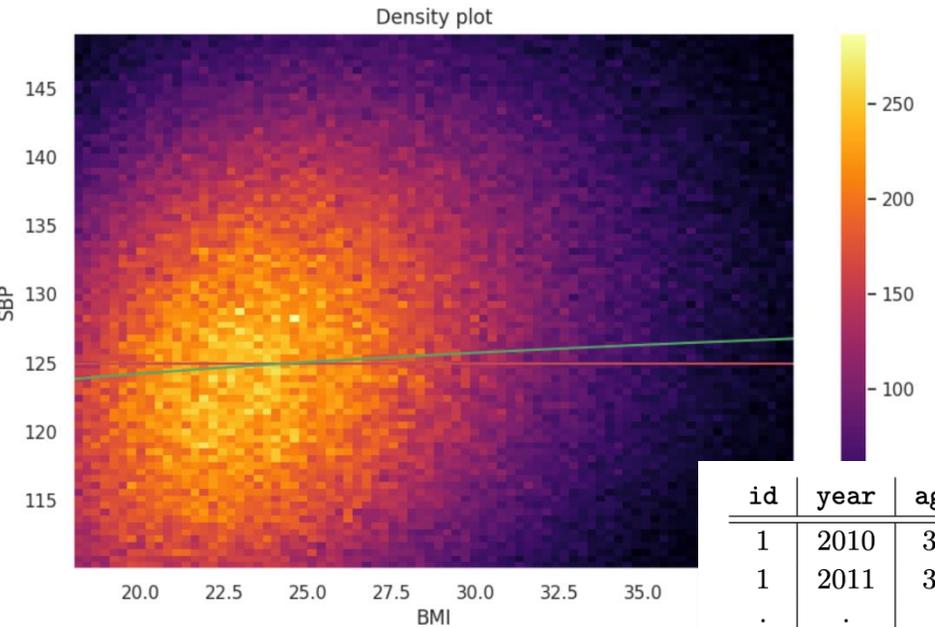
id	year	age	gender	bmi	sbp	event	time_to_event
1	2010	35	m	24	120	0	-
1	2011	36	m	24	120	0	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	2019	44	m	24	120	0	-
2	2010	35	m	33	145	0	7.5
2	2011	36	m	33	145	0	6.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	2017	44	m	33	145	1	0.5
3	2010	35	m	26	125	0	-
3	2011	36	m	26	125	0	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



# Creation of a synthetic health dataset

$$\begin{pmatrix} \text{sbp} \\ \log(\text{bmi}) \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 125 \\ 3.2 \end{pmatrix}, \begin{pmatrix} 15^2 & 15 \cdot 0.25\rho \\ 15 \cdot 0.25\rho & 0.25^2 \end{pmatrix} \right)$$

$$\mu^*(\mathbf{x}) = \exp \left( \log(\text{QRISK3}(\mathbf{x})) + r(\text{age}) \right) + 16(\text{num1} - 0.5)^4 + 4(\text{num2} - 0.5)^2 \text{num3} + \text{num3} + \text{binary} - 1.65$$



id	year	age	gender	bmi	sbp	event	time_to_event
1	2010	35	m	24	120	0	-
1	2011	36	m	24	120	0	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	2019	44	m	24	120	0	-
2	2010	35	m	33	145	0	7.5
2	2011	36	m	33	145	0	6.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	2017	44	m	33	145	1	0.5
3	2010	35	m	26	125	0	-
3	2011	36	m	26	125	0	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



## Model 1: Logistic regression/generalized linear model (GLM)

$$\mu_1(\mathbf{x}) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_1 - \dots - \beta_k x_k)}, \text{ or equivalently}$$

$$\text{logit}(\mu_1(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k,$$

```
import statsmodels.formula.api as sm
log_reg = sm.logit(formula="E~SBP+BMI+I(BMI**2)", data=time_to_event_train).fit()
pred = log_reg.predict(time_to_event_test)
```

### Odds/log-odds

$$\text{odds}(y = 1 \mid \mathbf{x}) := \frac{P(y = 1 \mid \mathbf{x})}{P(y = 0 \mid \mathbf{x})} = \frac{P(y = 1 \mid \mathbf{x})}{1 - P(y = 1 \mid \mathbf{x})}$$

$$\log(\text{odds}(y = 1 \mid \mathbf{x})) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

### Odds ratios

$$\frac{\text{odds}(y = 1 \mid (x_1, \dots, x_j + 1, \dots, x_k))}{\text{odds}(y = 1 \mid \mathbf{x})} = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_j (x_j + 1) + \dots + \beta_k x_k)}{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

$$= \exp(\beta_j).$$

id	year	age	gender	bmi	sbp	event	time_to_event
1	2010	35	m	24	120	0	-
1	2011	36	m	24	120	0	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	2019	44	m	24	120	0	-
2	2010	35	m	33	145	0	7.5
2	2011	36	m	33	145	0	6.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	2017	44	m	33	145	1	0.5
3	2010	35	m	26	125	0	-
3	2011	36	m	26	125	0	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

## Model 2: Cox regression

$$h(t | \mathbf{x}) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_k x_k)$$

```
import lifelines as ll
cph = ll.CoxPHFitter()
cph.fit(time_to_event, "T", event_col="E", formula="SBP+BMI+I(BMI**2)")
pred = (1 - np.array(cph.predict_survival_function(time_to_event_test))[10, :])
```

### From hazard rates to 10-year risk

$$\mu_2(\mathbf{x}) := 1 - \exp\left(-\int_0^{10} h(t | \mathbf{x}) dt\right)$$

### Hazard ratios

$$\frac{h_0(t) \exp(\beta_1 x_1 + \dots + \beta_j (x_j + 1) + \dots + \beta_k x_k)}{h_0(t) \exp(\beta_1 x_1 + \dots + \beta_k x_k)} = \exp(\beta_j)$$

id	year	age	gender	bmi	sbp	event	time_to_event
1	2010	35	m	24	120	0	-
1	2011	36	m	24	120	0	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	2019	44	m	24	120	0	-
2	2010	35	m	33	145	0	7.5
2	2011	36	m	33	145	0	6.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	2017	44	m	33	145	1	0.5
3	2010	35	m	26	125	0	-
3	2011	36	m	26	125	0	-
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



## Model performance

Table 3: Performance metrics on the test data subset of  $\mathcal{D}_1$ .

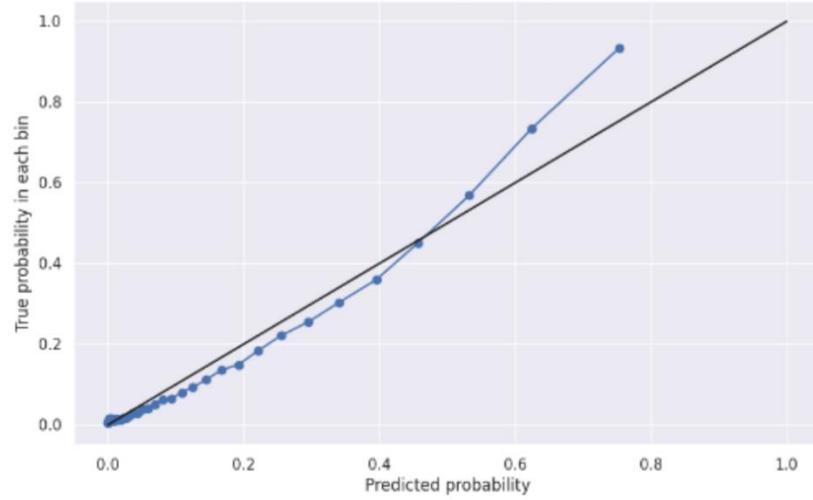
Performance metric	logistic regression $\mu_1(\mathbf{x})$	Cox regression $\mu_2(\mathbf{x})$	neural net $\mu_3(\mathbf{x})$
ROC AUC	56.17%	56.17%	56.04%
MSE wrt $\log(\mu^*(\mathbf{x}))$	0.0016	0.0016	0.0057
Logistic deviance	9223.88	9223.88	9227.72

Table 4: Performance metrics on the test data subset of  $\mathcal{D}_2$ .

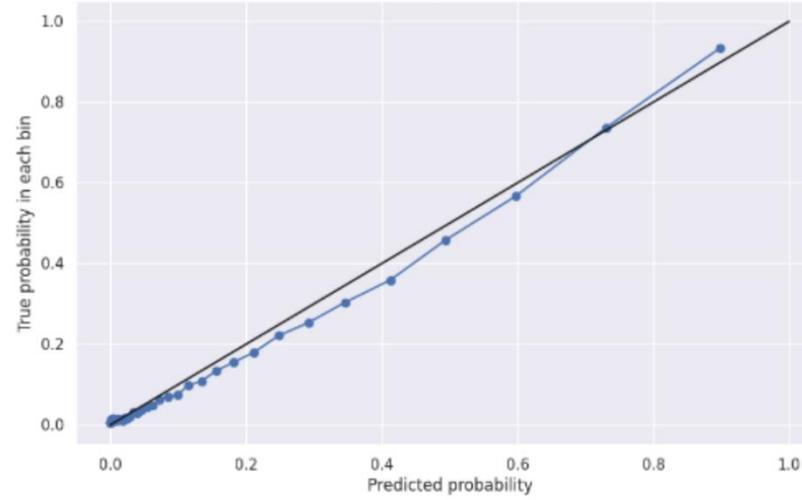
Performance metric	logistic regression $\mu_1(\mathbf{x})$	Cox regression $\mu_2(\mathbf{x})$	neural net $\mu_3(\mathbf{x})$
ROC AUC	90.54%	90.55%	92.05%
MSE wrt $\log(\mu^*(\mathbf{x}))$	1.74	1.75	0.11
Logistic deviance	85383	83994	75732

# Model performance

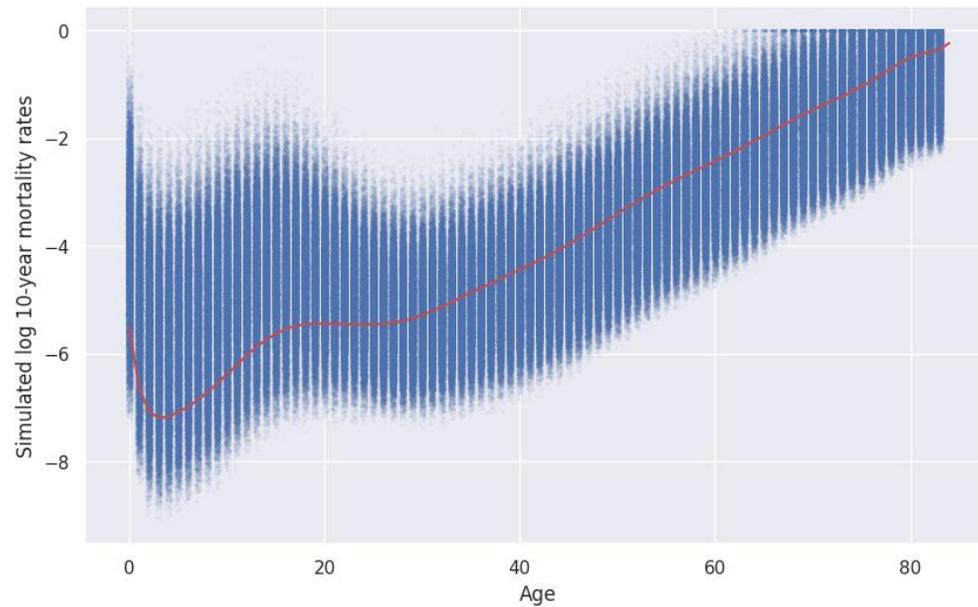
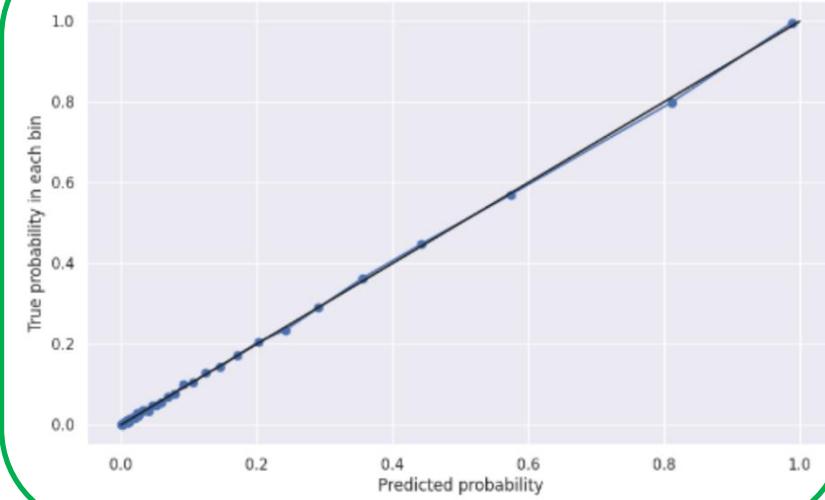
Calibration plot logistic regression



Calibration plot Cox regression



Calibration plot neural network



# Asymmetric cryptography (public/private key)

- Create a *shared secret*  $s$  for symmetric encryption (**stream ciphers**: Salsa20, RC4, ..., **block ciphers**: AES, DES, RC5, ...):

**Alice**: “Secret message”  $\rightarrow m \in (\mathbb{Z}/2\mathbb{Z})^n \rightarrow m + f(s)$  or  $f(m, s)$   $\Rightarrow$  **Bob**:  $m + f(s) + f(s)$  or  $f(f(m, s), s)$   $\rightarrow$  “Secret message”

- Some examples:

1. RSA (Rivest, Shamir, Adleman, 1977): **Factoring** large integers  $n = pq$  ( $n$  public,  $p, q$  private)
2. ElGamal (1985): **Discrete logarithm**, (multiplicative) group  $G$ , usually  $G \subset (\mathbb{Z}/p\mathbb{Z})^* =: \mathbb{F}_p^*$  of order  $q = (p-1)/2$  with generator  $g$ , solve  $x = \log_g h$  ( $g, G, h$  public,  $x$  private)
3. Elliptic curves methods (1985): **Discrete logarithm**, where group  $G$  is based on elliptic curves
4. Lattice based methods, e.g., LWE (“learning with errors”, 2005): Solve  $Ax + \varepsilon = b \pmod q$  for  $x \in (\mathbb{Z}/q\mathbb{Z})^n$ , where  $A$  is drawn uniformly from  $(\mathbb{Z}/q\mathbb{Z})^{m \times n}$ ,  $\varepsilon \in [-q/4, q/4]^m$  is drawn from a “non-trivial” distribution  $\chi$ , and  $b \in (\mathbb{Z}/q\mathbb{Z})^m$  ( $b, q, A$  public,  $x$  private)
5. Many more examples from [NIST standardization](#) proposals for post-quantum cryptography (factorization and discrete logarithm can be calculated very efficiently on quantum computers), e.g., CRYSTALS, 2018, while [SIKE](#) had to be removed from the list in August 2022.

# RSA

- [RSA, 1977](#) based on **Euler theorem**:  $m^{\varphi(n)} = 1 \pmod n$  for  $\gcd(m, n) = 1$ , where  $\varphi(n) = \#(\mathbb{Z}/n\mathbb{Z})^*$  (Euler's totient function)

–  $n = p = 7$ ,  $\varphi(p) = p - 1$

0	1	2	3	4	5	6
---	---	---	---	---	---	---

–  $n = pq = 15$ ,  $\varphi(pq) = (p - 1)(q - 1)$

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
---	---	---	---	---	---	---	---	---	---	----	----	----	----	----

– Choose “random”  $p, q, d$  with  $\gcd(d, \varphi(pq)) = 1$ , calculate  $e$  with  $ed = 1 \pmod{\varphi(pq)}$  with extended Euclidean algorithm,  $e, n$  public key,  $p, q, d$  private key

– Encryption: message  $m < n, m^e \pmod n$

– Decryption:  $m^{ed} = m \pmod n$

– *Proof*:  $m^{ed} = m^{k\varphi(n)+1} = m \pmod n$

– Calculating  $d$  from  $e$  and  $n \Leftrightarrow$  calculating  $\varphi(n) \Leftrightarrow$  **factoring**  $n = pq$

– *Proof idea*: “ $\Leftarrow$ ” 1.  $\varphi(pq) = (p - 1)(q - 1)$ , 2. extended Euclidean algorithm  $ed + b\varphi(pq) = \gcd(d, \varphi(pq)) = 1$

“ $\Rightarrow$ ” 1.  $\varphi(pq) = -(p + q) + 1 \pmod n$ , 2.  $ed - 1 = k\varphi(n)$  sufficient to factor  $n$  (see, e.g., [Miller, 1975](#), ERH)

– There are attacks for, e.g.,  $q < p < 2q, 3d < n^{1/4}$  ([Wiener, 1990](#)) and several others ([Zhang, 1999](#))

– Homomorphic encryption RSA example:  $(m_1 m_2)^e = m_1^e m_2^e \pmod n$ , in general:  $\text{enc}(\text{op}_1(m_1, m_2)) = \text{op}_2(\text{enc}(m_1), \text{enc}(m_2))$

## Key take-aways

1. Limited data availability for sensitive personal (Life & Health) data in practice (e.g., nFADP, 1 September 2023)
2. For sufficiently large and dense datasets, ML/DL methods outperform traditional models, creating value for policyholders and insurance companies
3. Privacy preserving methods can help to access data