What is a risk engine?
A fancy name for a prediction of risk from the characteristics of a person.
Let us look at one:
The UKPDS Risk Engine V2.

Download from: http://www.dtu.ox.ac.uk/

Two versions:
➢ UKPDS Risk Engine calculates risk one subject at a time,
➢ Excel version does it for a dataset.

The UKPDS Risk Engine V2
We need to put in the characteristics of the subject:
➢ Age, 60 years
➢ Duration of diabetes, 7 years
➢ Sex, Female
➢ Presence or absence of atrial fibrillation, Absent
➢ HbA1c, 7.0%
➢ Systolic blood pressure, 145 mm Hg
➢ Total cholesterol, 5.9 mmol/L
➢ HDL cholesterol, 1.1 mmol/L
➢ Ethnicity as White, Afro-Carribean, Asian-Indian,
➢ Smoking as Non-smoker, Ex-smoker, Current smoker.

We get the 10 year risk of Coronary Heart Disease, Fatal CHD, Stroke, Fatal Stroke.
The UKPDS Risk Engine V2

Characteristics of the subject:

Age = 60 years  
Duration of diabetes = 7 years
Sex = Female  
Atrial fibrillation = Absent
HbA1c = 7.0%  
SBP = 145 mm Hg
Total cholesterol = 5.9  
HDL cholesterol = 1.1
Ethnicity = White  
Smoking = Ex-smoker

10 year risks:

Coronary Heart Disease 14.0%
Fatal CHD 8.9%
Stroke 5.9%
Fatal Stroke 0.9%

The UKPDS Risk Engine V2

Characteristics of the subject:

Age = 70 years  
Duration of diabetes = 7 years
Sex = Female  
Atrial fibrillation = Absent
HbA1c = 7.0%  
SBP = 145 mm Hg
Total cholesterol = 5.9  
HDL cholesterol = 1.1
Ethnicity = White  
Smoking = Ex-smoker

10 year risks:

Coronary Heart Disease 14.0% ➔ 23.5%
Fatal CHD 8.9% ➔ 17.2%
Stroke 5.9% ➔ 13.5%
Fatal Stroke 0.9% ➔ 2.1%

The UKPDS Risk Engine V2

Characteristics of the subject:

Age = 70 years  
Duration of diabetes = 7 years
Sex = Female  
Atrial fibrillation = Present
HbA1c = 7.0%  
SBP = 145 mm Hg
Total cholesterol = 5.9  
HDL cholesterol = 1.1
Ethnicity = White  
Smoking = Ex-smoker

10 year risks:

Coronary Heart Disease 23.5% ➔ 23.5%
Fatal CHD 17.2% ➔ 17.2%
Stroke 13.5% ➔ 71.2%
Fatal Stroke 2.1% ➔ 11.2%
Where do these numbers come from?
The UK Prospective Diabetes Study (UKPDS)

Between 1977 and 1991, general practitioners in the catchment areas of 23 participating UKPDS hospitals were asked to refer all patients age 25 to 65 years presenting with newly diagnosed diabetes.

There were 5102 patients recruited to the study. Patients with a myocardial infarction within the last year, or with more than 1 vascular episode, were excluded.

Followed for between 6 and 20 years, median 11.4.

Record myocardial infarctions and strokes.


Where do these numbers come from?

We try to estimate the risk of CHD or stroke for diabetes subjects with any given combination of risk factors.

What is ‘risk’?

Risk is the probability that a person with particular characteristics will have the event, i.e. stroke, OR the proportion of people with these characteristics who will have the event.

What is a ‘risk factor’?

A risk factor is something which changes the risk. For example, smoking is a risk factor for stroke. Smokers have a greater risk than non-smokers.

Where do these numbers come from?

Characteristics of the subject:

Age = 60 years
Sex = ?
HbA1c = 7.0%
Total cholesterol = 5.9
Ethnicity = White
Smoking = Ex-smoker

10 year risks:

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronary Heart Disease</td>
<td>14.0%</td>
<td>19.8%</td>
</tr>
<tr>
<td>Fatal CHD</td>
<td>8.9%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Stroke</td>
<td>5.9%</td>
<td>8.3%</td>
</tr>
<tr>
<td>Fatal Stroke</td>
<td>0.9%</td>
<td>1.1%</td>
</tr>
</tbody>
</table>
Where do these numbers come from?

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronary Heart Disease</td>
<td>14.0%</td>
<td>19.8%</td>
</tr>
<tr>
<td>Fatal CHD</td>
<td>8.9%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Stroke</td>
<td>5.9%</td>
<td>8.3%</td>
</tr>
<tr>
<td>Fatal Stroke</td>
<td>0.9%</td>
<td>1.1%</td>
</tr>
</tbody>
</table>

Sex is a risk factor.

Men have a higher risk of CHD and of stroke than do women of the same age and with the same ethnicity, smoking history, blood pressure, etc.

Where do these numbers come from?

From the data, Stevens et al. estimated an equation for the risk of CHD:

Risk = mathematical function of age, sex, ethnicity, smoking, HBA1c, systolic BP, total cholesterol/HDL cholesterol ratio.

Why these variables?

Start with a set of variables known or suspected to be risk factors for CHD.

Are they significant predictors of risk in these data?

Put them into a regression model.

What is a ‘regression model’?

Aside on regression

What is a ‘regression model’?
Simple Linear Regression

Example: Muscle strength and height in 42 alcoholics

What is the relationship?
Regression: predict strength from observed height.

What is the mean strength for people with any given observed height?
Strength is the outcome, dependent, y, or left hand side variable.
Height is the predictor, explanatory, independent, x, or right hand side variable.
Linear relationship:
\[ \text{strength} = \text{intercept} + \text{slope} \times \text{height} \]
Equation of a straight line.

Simple Linear Regression

Which straight line should we choose?
Simple Linear Regression
Which straight line should we choose?

Choose the line which makes the distance from the points to the line \textit{in the y direction} a minimum.
Differences between the observed strength and the predicted strength.

Simple Linear Regression
Which straight line should we choose?

Minimise the sum of the squares of these differences.
Principle of least squares, least squares line or equation.

Simple Linear Regression
Can find confidence intervals and P values for the coefficients subject to assumptions.

Slope = 7.20, 95\% CI = 2.15 to 12.25, \(P=0.006\) against zero.
Intercept = –908, 95\% CI = –45 to –1771.
Multiple Linear Regression
More than one predictor:

Strength = –908 + 7.20 × height
Strength = 502 – 4.12 × age
Strength = –466 + 5.40 × height – 3.08 × age

Both coefficients are pulled towards zero because age and height are related:

Height = 179 – 0.195 × age,
P = 0.03
Age and height each explains some of the relationship between strength and the other.

Multiple Linear Regression
More than one predictor:

Strength = –908 + 7.20 × height
Strength = 502 – 4.12 × age
Strength = –466 + 5.40 × height – 3.08 × age

P = 0.04          P = 0.04

Compare:
Strength = –908 + 7.20 × height   P = 0.006
Strength = 502 – 4.12 × age       P = 0.007

Each predictor reduces the significance of the other because they are related to one another as well as to strength.
Why is it called regression?
Statistical method first used by Francis Galton (1886), who looked at predicting height of children from heights of parents. He found that if you choose tall parents, their children have a mean height less than their parents. If you choose short parents, their children have a mean height greater than their parents. He called this ‘regression towards mediocrity’. Now called ‘regression towards the mean’. Name ‘regression analysis’ stuck.

Other kinds of regression
Since Galton we have developed methods of regression analysis for other kinds of data, including:
- logistic regression for the probability of events,
- Cox regression for survival times,
plus many more.
Stevens et al. used a combination of logistic and Cox regression. Their formula looks fiendishly complicated.

Back to ‘where do these numbers come from?’
From the data, Stevens et al. estimated an equation for the risk of CHD:
Risk = mathematical function of age, sex, ethnicity, smoking, HBA1c, systolic BP, total cholesterol/HDL cholesterol ratio.
Why these variables?
Start with a set of variables known or suspected to be risk factors for CHD.
Are they significant predictors of risk in these data?
Put them into a regression model.
Drop variables which do not have a significant predicting effect.
Where do these numbers come from?
From the data, Stevens et al. estimated an equation for the risk of CHD:
Risk = mathematical function of age, sex, ethnicity, smoking, HBA1c, systolic BP, total cholesterol/HDL cholesterol ratio.
Drop variables which do not have a significant predicting effect.
Example: serum triglyceride predicts risk, but not when we include cholesterol as well.
We do not need triglyceride if we have cholesterol.

Where do these numbers come from?
The UKPDS regression equation:
Risk of CHD event in \( t \) years in a patient with diabetes diagnosed \( T \) years ago at age \( \text{AGE} \) is \( R(t) \) where
\[
R(t) = 1 - \exp\left(-q \frac{d'T (1-d')}{1-d}\right)
\]
\( d = 1.078 \) and \( q \) is given by
\[
q = 0.0112 \times 1.059^{\text{AGE} - 55} \times 0.525^{\text{SEX}} \times 0.390^{\text{AFROCARIB}} \times 1.350^{\text{SMOKER}} \times 1.183^{\text{HBA1c}} - 0.72 \times 1.088^{0.375 \times \text{SBP} - 135.7} \times 3.845^{0.3 \times \text{LR} - 1.59}
\]
(LR = lipid ratio, total cholesterol/HDL cholesterol.)
Which is why they wrote the free program.

Confidence intervals
Risk is presented as follows:

<table>
<thead>
<tr>
<th>0</th>
<th>15</th>
<th>30</th>
<th>100</th>
</tr>
</thead>
</table>

 Shows the point estimate and 95% confidence interval.
Point estimate = best estimate of the proportion of people with these characteristics who will experience event.
Confidence interval = range of values within which we think proportion should lie.
Not the error for the individual risk of person.
Confidence intervals

Risk is presented as follows:

Means that our best estimate that 27% of people with these characteristics will have CHD event in next 10 years.
Might be more, might be less, but if it were not between 22% and 32%, the UKPDS sample would be unusual.

Confidence intervals

The main purpose of these intervals is to warn us when the model is very imprecise. This happens when subjects would be at the extremes of some of the distributions in the UKPDS sample.

For example, a man aged 93, diabetic for 45 years, with HbA1c 9.0%, SBP 178, total cholesterol 9.0, HDL cholesterol 1.0, with atrial fibrillation.

We get very wide confidence intervals, because this is a very extreme individual.

Regression dilution

Characteristics of the subject:

- Age = 60 years
- Sex = Female
- HbA1c = 7.0%
- Total cholesterol = 5.9
- Ethnicity = White
- Smoking = Ex-smoker
- Duration of diabetes = 7 years
- Atrial fibrillation = Absent
- SBP = 145 mm Hg — mean of 3
- HDL cholesterol = 1.1

10 year risks:
- Coronary Heart Disease: 14.0% → 14.4%
- Fatal CHD: 8.9% → 9.1%
- Stroke: 5.9% → 6.1%
- Fatal Stroke: 0.9% → 1.0%

Adjusted for regression dilution
Regression dilution
If we tell the risk engine that systolic blood pressure is the mean of four measurements, we get different estimates. We also get a message:

Adjusted for regression dilution

What is regression dilution?

What is regression dilution?
An example: we have the true values of the outcome and predictor and predictor predicts outcome exactly.

No error in either variable.
Regression line goes exactly through points.

What is regression dilution?
An example: we have the true values of the outcome and predictor and predictor predicts outcome exactly.

Error in outcome variable.
Regression line is unchanged, apart from a bit of random variation.
What is regression dilution?

An example: we have the true values of the outcome and predictor and predictor predicts outcome exactly.

Error in predictor variable.

Regression line is changed, slope closer to zero.

Regression predicts mean outcome from observed predictor.

What is regression dilution?

An example: we have the true values of the outcome and predictor and predictor predicts outcome exactly.

Error in predictor variable, but predictor is mean of four observations. Error is halved.

Regression line is changed, slope is between slope for full error and slope for no error.

What is regression dilution?

An example: we have the true values of the outcome and predictor and predictor predicts outcome exactly.

Regression predicts mean outcome from observed predictor.

If we have a regression for a single observation with error, then put in the mean of several observations, prediction will be wrong.
Are there other cardiac risk calculators?
Yes.
Joint British Societies Cardiac Risk Assessor
http://www.bnf.org/bnf/extra/50/noframes/450024.htm
This requires similar but not identical data.
Treats diabetes as yes or no.
Based on the Framingham study, Massachusetts, USA.

<table>
<thead>
<tr>
<th></th>
<th>UKPDS</th>
<th>JBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>60 years</td>
<td>60 years</td>
</tr>
<tr>
<td>Duration of diabetes</td>
<td>7 years</td>
<td>No</td>
</tr>
<tr>
<td>Presence of diabetes</td>
<td>Assumed</td>
<td>Yes</td>
</tr>
<tr>
<td>Sex</td>
<td>Female</td>
<td>Female</td>
</tr>
<tr>
<td>Atrial fibrillation,</td>
<td>Absent</td>
<td>No</td>
</tr>
<tr>
<td>ECG-LVH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HbA1c</td>
<td>7.0%</td>
<td>7.0%</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>145 mm Hg</td>
<td>145 mm Hg</td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td>85 mm Hg</td>
<td>85 mm Hg</td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>5.9 mmol/L</td>
<td>5.9 mmol/L</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>1.1 mmol/L</td>
<td>1.1 mmol/L</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>White</td>
<td>Ex-smoker</td>
</tr>
<tr>
<td>Smoking</td>
<td>Ex-smoker</td>
<td>Non-smoker</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>UKPDS</th>
<th>JBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD event</td>
<td>14.0%</td>
<td>20.5% (20.9%)</td>
</tr>
<tr>
<td>Fatal CHD</td>
<td>8.9%</td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>5.9%</td>
<td>5.9% (5.7%)</td>
</tr>
<tr>
<td>Fatal stroke</td>
<td>0.9%</td>
<td></td>
</tr>
<tr>
<td>Increase HbA1c from 7.0% to 10.0%:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UKPDS</td>
<td>JBS</td>
</tr>
<tr>
<td>CHD event</td>
<td>20.3%</td>
<td>20.5% (20.9%)</td>
</tr>
<tr>
<td>Fatal CHD</td>
<td>14.6%</td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>5.9%</td>
<td>5.9% (5.7%)</td>
</tr>
<tr>
<td>Fatal stroke</td>
<td>0.9%</td>
<td></td>
</tr>
</tbody>
</table>
### Problems with risk calculators

Apply only to the population from which original data drawn.

UKPDS participating centres: Radcliffe Infirmary, Oxford; Royal Infirmary, Aberdeen; University Hospital, Birmingham; St George's Hospital and Hammersmith Hospital, London; City Hospital, Belfast; North Staffordshire Royal Infirmary, Stoke-on-Trent; Royal Victoria Hospital, Belfast; St Helier Hospital, Carshalton; Whittington Hospital, London; Norfolk and Norwich Hospital; Lister Hospital, Stevenage; Ipswich Hospital; Ninewells Hospital, Dundee; Northampton Hospital; Torbay Hospital; Peterborough General Hospital; Scarborough Hospital; Derbyshire Royal Infirmary; Manchester Royal Infirmary; Hope Hospital, Salford; Leicester General Hospital; Royal Devon and Exeter Hospital.

Not York!

---

### Problems with risk calculators

Apply only to the population from which original data drawn.

Our new subject may differ in several ways:

- **Geographical**: risks may vary due to environment, e.g. hard/soft water, effectiveness of local health service.
- **Time**: risks may have changed since data collected, e.g. APACHE score gives risks which are too high.
- **Ethnic group**: may be ethnic factors in risk, e.g. due to diet, genetics, alcohol consumption, etc.
- **Other factors**: Factors unimportant in the original data may be important in our population, e.g. predicting cot death.

A Chinese origin subject living in York?

---

### Problems with risk calculators

Apply only to the population from which original data drawn.

UKPDS applies only to type II diabetes, not to type I.

Framingham does not distinguish between type I and type II.

Framingham will not have had a great number of diabetics in the original sample, so may not be very accurate for them.
Problems with risk calculators
No way of checking.
If risk is 20% over 10 years and the subject has a CHD event, was the risk estimate right or wrong?
Can’t say: it was a probability.
Need many subjects.

Problems with risk calculators
Based on observational studies not trials.
We do not know whether, say, stopping smoking would have the effect the risk engine would suggest.
Cannot say whether a reduction in cholesterol produced by statins would have the same effect as one produced by diet, because of possible effects on other variables.