Using Electronic Health Records to Predict Health Risks Associated with Obesity

Volker Schnecke
Data Science – Global Development
Novo Nordisk
Who are we?

- Novo Nordisk: global pharmaceutical company headquartered in Denmark
- Novo Nordisk develops, manufactures and markets drugs to treat diabetes, obesity, haemophilia, and growth hormone disorders
- Pharmaceutical industry is a data-driven business
- We are part of Global Development, i.e., the function in R&D that runs clinical trials to demonstrate that our drugs are safe and effective
- We are 3 big data analysts (data scientists), 5 data science programmers (data engineers), and 1 manager
Clinical trials

- Homogenous population
- Treatment group randomly chosen
- Data sampling at fixed time points
- All relevant measurements for all patients (dense data)
- Demonstrate drug efficacy in controlled clinical environment

Observational data

- Heterogenous population
- Treatment group predefined
- Data sampling at random time points
- Some relevant measurements for some patients (sparse data)
- Demonstrate drug efficacy in real-world scenario

Electronic Health/Medical Records capture observational data
Obesity is a chronic disease in which excess body fat has accumulated to an extent that it may have a negative health effect.

Defined via Body Mass Index (BMI):

\[
\text{BMI} = \frac{\text{weight}[\text{kg}]}{\text{height}[\text{m}] \times \text{height}[\text{m}]} 
\]

- Normal weight: 18.5 ≤ BMI < 25
- Overweight: 25 ≤ BMI < 30
- Obesity: BMI ≥ 30

England\(^1\): 40% of men and 31% of women are overweight;
27% of men and 30% of women have obesity.

Persons with BMI 40-45 have a **8-10 years shorter life expectancy** than persons with BMI 22.5-25 (data from North America and Western Europe)\(^2\).

Sources:
- \(^1\) Health Survey England (HSE), 2017
- \(^2\) Lancet, 373 (9669), pp.1038-1096, 2009
BMI (Body mass index)

- BMI is easy to measure / calculate
- No differentiation between muscle and fat mass
- Additional, more informative measures:
  - waist circumference
  - waist-hip ratio
  - waist-to-height ratio

<table>
<thead>
<tr>
<th>BMI</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>72kg</td>
</tr>
<tr>
<td>25</td>
<td>90kg</td>
</tr>
<tr>
<td>30</td>
<td>108kg</td>
</tr>
<tr>
<td>35</td>
<td>126kg</td>
</tr>
<tr>
<td>40</td>
<td>144kg</td>
</tr>
<tr>
<td>45</td>
<td>163kg</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9m</td>
<td>1.8m</td>
<td></td>
</tr>
<tr>
<td>1.91m</td>
<td>1.86m</td>
<td></td>
</tr>
<tr>
<td>1.65m</td>
<td>1.68m</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Height</th>
<th>Weight</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Emma Watson</td>
<td>1.65m</td>
<td>54kg</td>
</tr>
<tr>
<td>2</td>
<td>Arnold Schwarzenegger (age 30)</td>
<td>1.86m</td>
<td>104kg</td>
</tr>
<tr>
<td>3</td>
<td>Donald Trump</td>
<td>1.91m</td>
<td>110kg</td>
</tr>
<tr>
<td>4</td>
<td>Vladimir Putin</td>
<td>1.68m</td>
<td>71kg</td>
</tr>
</tbody>
</table>

Sources: healthyceleb.com, bodywhat.com
What are the risks of having obesity?

- What are the most likely health challenges a person with obesity will face in the future?
- There are metabolically healthy people with obesity
- Is there an obesity paradox? For example, a BMI of 25 has the lowest association with death[1]
- Can we quantify the risk of future diseases or serious health events for a given individual?

[1] Lancet Diabetes Endocrinol, 6, pp. 944-953, 2018
Risk assessment of CV death

- Provides estimate of 10-year risk of fatal cardiovascular (CV) event, i.e., death by stroke or heart attack
- Risk estimate based on gender, age, total cholesterol, blood pressure and smoking
- Used for treatment guidelines
- Used for enriching clinical trials with high-risk patients
Electronical Medical Records

- Purpose is to capture doctor’s notes, patients’ history, allergies, drug prescriptions, results from blood or urine tests

- Anonymized by vendors and often distributed as flat text files (dump of relational database tables)
Extracting the relevant data

Comorbidities

Baseline period

Future events

**T2D**: Type-2 diabetes  
**MI**: Myocardial infarction  
**HF**: Heart failure  
**CKD**: Chronic kidney disease

---|---|---|---|---|---|---|---|---
**T2D** | **BMI** | Female  | Age 35  | BMI 32  | Non-smoker | **MI** | **CKD** | **HF**

**BMI**  
Male  
Age 42  
BMI 39  
Smoker

**BMI**  
Male  
Age 63  
BMI 22  
Non-smoker  
2004

**BMI**  
Female  
Age 35  
BMI 32  
Non-smoker

**High cholesterol**  
**T2D**  
**Asthma**  
**Hypertension**  
**Stroke**

**Comorbidities**

**Baseline period**

**Future events**

**T2D**: Type-2 diabetes  
**MI**: Myocardial infarction  
**HF**: Heart failure  
**CKD**: Chronic kidney disease
Median age: 51 years
Median follow-up period: 11 years

Diagnoses/events during follow-up:
- Chronic kidney disease: 9.0%
- Type-2 diabetes: 5.9%
- Heart failure: 3.0%
- Stroke: 2.3%

Categorized by BMI

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>1,061,481</td>
</tr>
<tr>
<td>25-30</td>
<td>1,038,409</td>
</tr>
<tr>
<td>30-35</td>
<td>489,855</td>
</tr>
<tr>
<td>35-40</td>
<td>170,156</td>
</tr>
<tr>
<td>40-45</td>
<td>64,864</td>
</tr>
</tbody>
</table>
Kaplan-Meier plots based on population of 2.8 million UK residents with median follow-up time of 11 years

Prediction of 10-year risk for 13 outcomes (diagnoses/event):
• Separate Cox proportional hazard model for each outcome
• Age as underlying time variable
• Covariates: BMI, gender, smoking status, 12 comorbidities
Obesity raises the risk of early death by up to 50%
Results from Cox proportional hazard models based on 2.8 million UK residents with age as underlying time variable, adjusted for gender, smoking, and diagnoses of 12 comorbidities at baseline.
Can we make more accurate predictions?

Additional data in Electronic Medical Records:

- **Drug prescriptions:**
  - Drug brand name
  - Drug class
  - Sometimes package size and dose

- **Vitals:**
  - Blood pressure
  - Heart rate

- **Biomarker values:**
  - Results of laboratory tests, e.g., of blood or urine samples
  - For example cholesterol level or white blood cell count
Better prediction by including biomarkers

“Healthy”:
- Hypertension
- T2D

“Heart failure”:
- Hypertension
- T2D
- Stroke
- LDL
- HDL
- TG

Classification:
- Male
- Age 51
- BMI 26
- Hypertension
- Dyslipidaemia
- LDL 5.5 mmol/L
- HDL 2.3 mmol/L

Model:
- Healthy
- Heart failure

T2D: Type-2 diabetes
AF: Atrial fibrillation
TG: Triglycerides
LDL: LDL cholesterol
HDL: HDL cholesterol

Male
Age 42
BMI 31
Hypertension
Atrial fibrillation

Male
Age 42
BMI 28
Hypertension
Type-2 diabetes
Stroke

Male
Age 62
BMI 28
Hypertension
Type-2 diabetes
Stroke

Triglycerides 3.8 mmol/L
HDL 2.3 mmol/L

LDL 5.7 mmol/L
HDL 1.9 mmol/L
Triglycerides 4.6 mmol/L
**Study cohort**

Mean number of biomarker values per patient: 16.5 (of 38)

Data source: GE centricity, EMRs of US residents
Simple models: BMI, age, gender, blood pressure, heart rate + 12 comorbidities

Advanced models: also levels of biomarkers measured during last year

Data from approximately 1,100,000 individuals, predictions from gradient boosting models (XGBoost) on external test set of 20% of study cohort.
Model usage for clinical trial design

- Two-year clinical trial to test new anti heart failure (HF) drug
- 2% of population will get HF during 2 years
- 400 events needed to show drug efficacy
- 20,000 patients need to be included in clinical trial

<table>
<thead>
<tr>
<th></th>
<th>Healthy</th>
<th>HF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predicted Healthy</strong></td>
<td>19,608</td>
<td>88</td>
</tr>
<tr>
<td><strong>Predicted HF</strong></td>
<td>4,304</td>
<td>400</td>
</tr>
</tbody>
</table>

- 82% accuracy to predict both groups
- Predicting 24,400 patients will identify 400 “likely HF patients”
- 4,704 patients need to be included in trial
Stumbling blocks

- Bias in study cohort:
  - Requirement of more biomarkers results in older, more diseased population

- Distinguishing interesting (very high) biomarker values from typos

- Skewed class distributions
• Data not collected for our purpose ("secondary use")
• No control over data collection
• Data is of larger scale than in-house data, i.e., IT infrastructure needed to be built ‘from scratch’
• Not really big data, but big enough to (sometimes) benefit from using big-data tools (Hive, Spark, SparklyR)
• Workflows typically involve different platforms (EMR cluster, Hive database, local database, local Rstudio)
• Different workflows for different data sources
Data visualization interfaces

Web interface for visualizing key data in patient’s journal style
Stakeholder management is key

- Often too high / too low expectations on what we can do with our data
- Most of us are (fairly) new to the data sources, so often need to do feasibility analyses before responding or committing to a stakeholder’s request

- Coordination of activities / competence development with data science teams in other parts of the organization

- Our vision: “We want to be the ones to provide answers to questions that our stakeholders didn’t know they could ask”
We’re living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you’re often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we’re uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for the New York Times as well as a national security and political contributor for CNN and a frequent guest on CBS This Morning, Face the Nation, and many PBS shows.