



# LEVERAGING DATA TO PREVENT & MANAGE DIABETES

Luke Slawomirski  
Consultant\* - OECD - Directorate for Employment, Labour & Social Affairs

IDF Europe World Diabetes Day Symposium  
16 Nov 2021

*\* Content reflects presenter's opinions*



## First of all ...

---





# Many industries have transformed themselves around digital technology (data = lifeblood)\*



- Better products
- Better services
- More efficient & productive
- Big consumer surpluses

\*re-use for 2<sup>o</sup> purposes



# The possibilities (in health) are tantalising...

**Open access**

**BMJ Open** Google search history presenting to an emergency department: an observational study

Jeremy M Asch,<sup>1,2</sup> David A Asch,<sup>1,2</sup> Elesa Norah Sadek,<sup>1</sup> Raina M Merchant<sup>1,2</sup>

**ABSTRACT**  
**Objective:** To test patients' willingness to share and link their prior Google search history with data from their electronic medical record (EMR), and to explore associations between search histories and clinical conditions.  
**Design:** Cross-sectional study of emergency department (ED) patients from 2016 to 2017.  
**Setting:** Academic medical centre ED.  
**Participants:** A total of 700 patients were approached, 534 of a random sample of 611 (91%) provided a link to a Google account; 165 of these (30%) consented to share their Google search histories and EMR data; 119 (72%) were able to do so. 19 (13%) of these 119 patients had no data and were not included in the final count. Patients under the age of 18 or with a triage level of 1 were considered ineligible and were not approached.  
**Main results:** 100% search histories of searches in the remote past and within 7 days of the ED visit, and associations between patients' clinical and demographic characteristics and their internet search volume and search content.  
**Results:** The 103 participants yielded 551 421 unique search queries; 37 418 (7%) were health related in the 7 days prior to an ED visit, the percentage of health-related searches was 15%. During that time, 50% of queries searched for symptoms, 53% for information about a hospital and 23% about the treatment or management of a disease. 53% of participants who used Google in the week leading up to their ED visit searched for content directly related to their chief complaint.  
**Conclusions:** Patients were willing to allow researchers substantial access to their Google search histories and their EMR data. This change in volume and content of search activity prior to an ED visit suggests opportunities to anticipate and improve health care utilization in advance of ED visits.  
 Digital media capture and document an increasing segment of our personal lives in the tracks left on online or in-store purchases, wearable devices or engagement with social media. Many of these digital traces reflect health. Facebook, Twitter and Instagram posts can reveal health-related behaviours, symptoms or diagnoses.<sup>1-3</sup> But while these social media posts reflect what

Teaser text for the paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2018-024791>).

Received 12 June 2018  
 Revised 14 November 2018  
 Accepted 21 December 2018

Check for updates

© Author(s) 2019. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

Department of Emergency Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA  
 Center for Digital Health, Penn Medicine, Philadelphia, Pennsylvania, USA  
 The Center for Health Equity Research and Promotion, Philadelphia Veterans Affairs Medical Center, Philadelphia, Pennsylvania, USA

Correspondence to: Jeremy M Asch; [jeremy.asch@upenn.edu](mailto:jeremy.asch@upenn.edu)

BMJ 2019;399:e024791. doi:10.1136/bmjopen-2018-024791

**SCIENCE REPORTS**

**OPEN** Early detection of type 2 diabetes mellitus using machine learning-based prediction models

Leon Kopitar<sup>1,2,3</sup>, Primoz Kocbek<sup>1,2</sup>, Leona Cilar<sup>1</sup>, Aziz Sheikh<sup>4,5</sup> & Gregor S. ...

**ABSTRACT**  
 Most screening tests for T2DM in use today were developed using multivariate regression that are often further simplified to allow transformation into a scoring formula. The of electronically collected data opened the opportunity to develop more complex, machine learning-based prediction models (i.e. Gbmnet, RF, XGBoost, LightGBM) to regression models for prediction of undiagnosed T2DM. The performance in predicting plasma glucose level was measured using 100 bootstrap iterations in different subsets simulating new incoming data in 5-month batches. With 5 months of data available model performed with the lowest average RMSE of 0.838, followed by RF (0.842), LightGBM (0.855) and XGBoost (0.881). When more data were added, Gbmnet improved rate (+2.4%). The highest level of variable selection stability over time was observed models. Our results show no clinically relevant improvement when more sophisticated models were used. Since higher stability of selected variables over time contributes interpretation of the models, interpretability and model calibration should also be development of clinical prediction models.  
 Type 2 diabetes mellitus (T2DM) is very common and is responsible for very considerable Furthermore, it is a substantial financial drain both on individuals/families, health systems major concern is that the incidence and prevalence of T2DM are increasing rapidly – it is estimated that 425 million people had any type of diabetes (approx. 5.5% of worldwide population) had T2DM and according to projection estimations the prevalence is going to increase substantially, for example, a 48% increase of prevalence from the above numbers is estimated to be 629 million people (approx. 6.6% of the worldwide population) are estimated to be from any type of diabetes.<sup>1</sup> T2DM can lead to substantially increased risk of macrovascular disease, especially in those with inadequate glycaemic control.<sup>2</sup> Progression of T2DM to glucose is typically slow and more importantly, its symptoms may remain undetected for a diagnosis are an important contributory factor to poor control and risk of complications.<sup>3</sup>  
 Data mining is nowadays applied to various fields of science, including healthcare and is pattern recognition, disease prediction and classification using various data mining to increased prevalence of T2DM, various techniques have been used to build predictive models for early disease diagnosis, such as logistic and Cox proportional hazard regression models<sup>4</sup> based on clinical data<sup>5,6</sup>, etc. The study by Hansen et al.<sup>7</sup> showed that logistic regression (n = 365) models for risk estimation in the general population. Even though there are multiple to build prediction models, prediction accuracy and data validity are often not realistic for practice. Models also perform well in specific dataset where they were developed but are not adapt sufficiently well with used in other datasets.<sup>8</sup>

**KEYWORDS:** (type 2) diabetes mellitus, machine learning, prediction models, type 2 diabetes mellitus

1/23

10.1038/s41598-019-68771-4

Lai et al. BMC Endocrine Disorders (2019) 19:101  
<https://doi.org/10.1186/s12902-019-0436-6>

BMC Endocrine Disorders

**RESEARCH ARTICLE** Open Access

**Predictive models for diabetes mellitus using machine learning techniques**

Hang Lai<sup>1,2</sup>, Huaxiong Huang<sup>1,2</sup>, Karim Keshavjee<sup>3,2</sup>, Aziz Guergachi<sup>3,2,4</sup> and Xin Gao<sup>1,2\*</sup>

**Abstract**  
**Background:** Diabetes Mellitus is an increasingly prevalent chronic disease characterized by the body's inability to metabolize glucose. The objective of this study was to build an effective predictive model with high sensitivity and selectivity to better identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic data and the laboratory results during their visits to medical facilities.  
**Methods:** Using the most recent records of 13,300 Canadian patients aged between 18 and 90 years, along with their laboratory information (age, sex, fasting blood glucose, body mass index, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein), we built predictive models using Logistic Regression and Gradient Boosting Machine (GBM) techniques. The area under the receiver operating characteristic curve (AUROC) was used to evaluate the discriminatory capability of these models. We used the adjusted threshold method and the class weight method to improve sensitivity – the proportion of Diabetes Mellitus patients correctly predicted by the model. We also compared these models to other learning machine techniques such as Decision Tree and Random Forest.  
**Results:** The AUROC for the proposed GBM model is 84.7% with a sensitivity of 71.6% and the AUROC for the proposed Logistic Regression model is 84.0% with a sensitivity of 73.4%. The GBM and Logistic Regression models perform better than the Random Forest and Decision Tree models.  
**Conclusions:** The ability of our model to predict patients with Diabetes using some commonly used lab results is high with satisfactory sensitivity. These models can be built into an online computer program to help physicians in predicting patients with future occurrence of diabetes and providing necessary preventive interventions. The model is developed and validated on the Canadian population which is more specific and powerful to apply on Canadian patients than existing models developed from US or other populations. Fasting blood glucose, body mass index, high-density lipoprotein, and triglycerides were the most important predictors in these models.  
**Keywords:** Diabetes mellitus, Machine learning, Gradient boosting machine, Predictive models, Misclassification cost

**Background**  
 Diabetes Mellitus (DM) is an increasingly prevalent chronic disease characterized by the body's inability to metabolize glucose. Finding the disease at the early stage helps reduce medical costs and the risk of patients having more complicated health problems. Wilson et al. [1] developed the Framingham Diabetes Risk Scoring Model (FDRSM) to predict the risk for developing DM in middle-aged American adults (45 to 64 years of age) using Logistic Regression. The risk factors considered in this simple clinical model are parental history of DM, obesity, high blood pressure, low levels of high-density lipoprotein cholesterol, elevated triglyceride levels, and impaired fasting glucose. The number of subjects in the sample was 3140 and the area under the receiver operating characteristic curve (AUROC) was reported to be 85.0%. The performance of this algorithm was evaluated in a Canadian population by Mashayekhi et al. [11] using the same predictors as Wilson et al. [1] with the

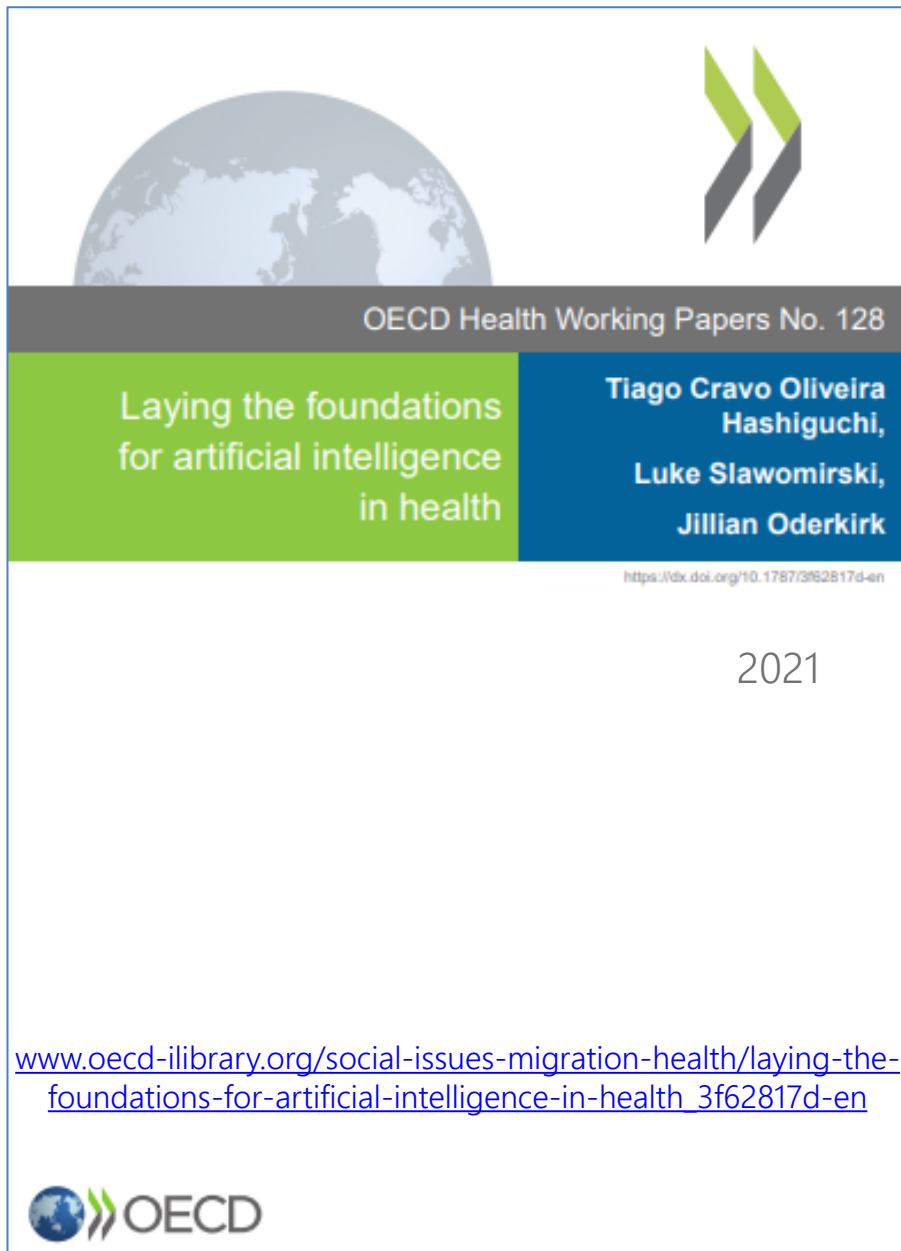
\* Correspondence: [hanglai@stat.yorku.ca](mailto:hanglai@stat.yorku.ca)  
 Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada  
 The Fields Institute for Research in Mathematical Sciences, Center for Quantitative Analysis and Modelling (CQAM) Lab, 232 College Street, Toronto, Ontario M5T 2E1, Canada  
 Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

- ✓ Prediction
- ✓ Detection
- ✓ Prevention
- ✓ Clinical Rx
- ✓ Policy Rx

BUT...





"Policy makers should beware the hype of AI\* in health care

...

In setting the foundations for AI to help achieve health policy objectives, one key priority is to **improve data quality, interoperability and access** in a secure way through better data governance."

*\*machine learning -- most often probabilistic models (regression / curve fitting anyway)*

# Health systems

---

“data rich ... information poor”





# Countries routinely linking health, contextual and outcomes data (for 2<sup>o</sup> purposes)

Multiple health care settings	AUS	BEL	CAN	DEN	FIN	FRA	ISR	KOR	LVA	NLD	NOR	SGP	SVN	USA	14
Disease registries to mortality data	AUS	AUT	CAN	CZE	EST	FIN	ISR	JPN	KOR	LVA	LUX	SGP	SVN	SWE	14
Health care to mortality data	AUS	AUT	CZE	FIN	ISR	KOR	LVA	NLD	SVN	USA					10
Population census to disease registry data	CAN	EST	FIN	LVA	NLD	SVN	SWE								7
Population census to health care data	AUS	CAN	FIN	NLD	SVN	SWE									6
Cancer registry to health care data	BEL	CZE	LVA	NLD	SWE										5
Population health survey to health care data	FIN	SWE	USA												3
Cancer registry to cancer screening data	LUX	SVN													2
Health care to personal income tax data	AUS	NLD													2
Population health survey to social insurance data (social security)	SWE	USA													2
Population health survey to Population census data	NLD	SWE													2

23 OECD countries





# ...using clinical (EMR) data ...

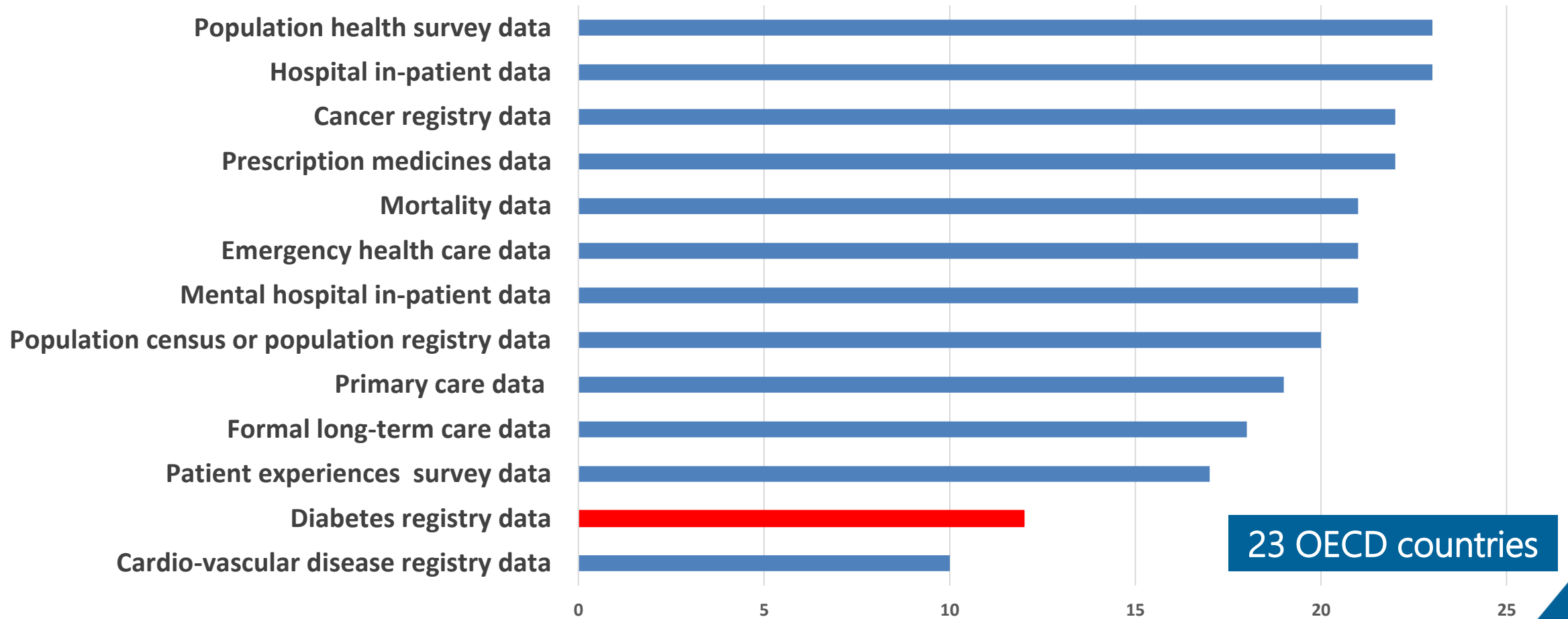
[www.oecd-ilibrary.org/social-issues-migration-health/survey-results-national-health-data-infrastructure-and-governance\\_55d24b5d-en](http://www.oecd-ilibrary.org/social-issues-migration-health/survey-results-national-health-data-infrastructure-and-governance_55d24b5d-en)

23 OECD countries

16						
BEL						
CRI						
DEN						
FIN	12					
NLD	BEL					
SWE	CRI	10	10			
TUR	DEN	BEL	BEL			
HUN	FIN	CRI	CRI	8		
ISL	NLD	DEN	DEN	CRI	7	
ISR	SWE	FIN	DEU	DEN	CRI	6
LTU	TUR	NLD	FIN	FIN	DEN	BEL
CZE	ISL	SWE	NLD	ISR	ISR	EST
EST	ISR	TUR	SWE	LUX	LUX	DEU
JPN	LTU	LTU	TUR	NLD	NLD	ISR
PRT	EST	CZE	ISL	PRT	PRT	ITA
SLO	JPN	PRT	ISR	TUR	SWE	NLD
Public health monitoring	Monitoring patient safety	Monitoring health system performance	Medical and health care research	Data Mining to find/extract data within EHRs	Predictive analytics trained on EHRs	Linkage of EHRs and genomic, environmental, behavioural, economic or other data

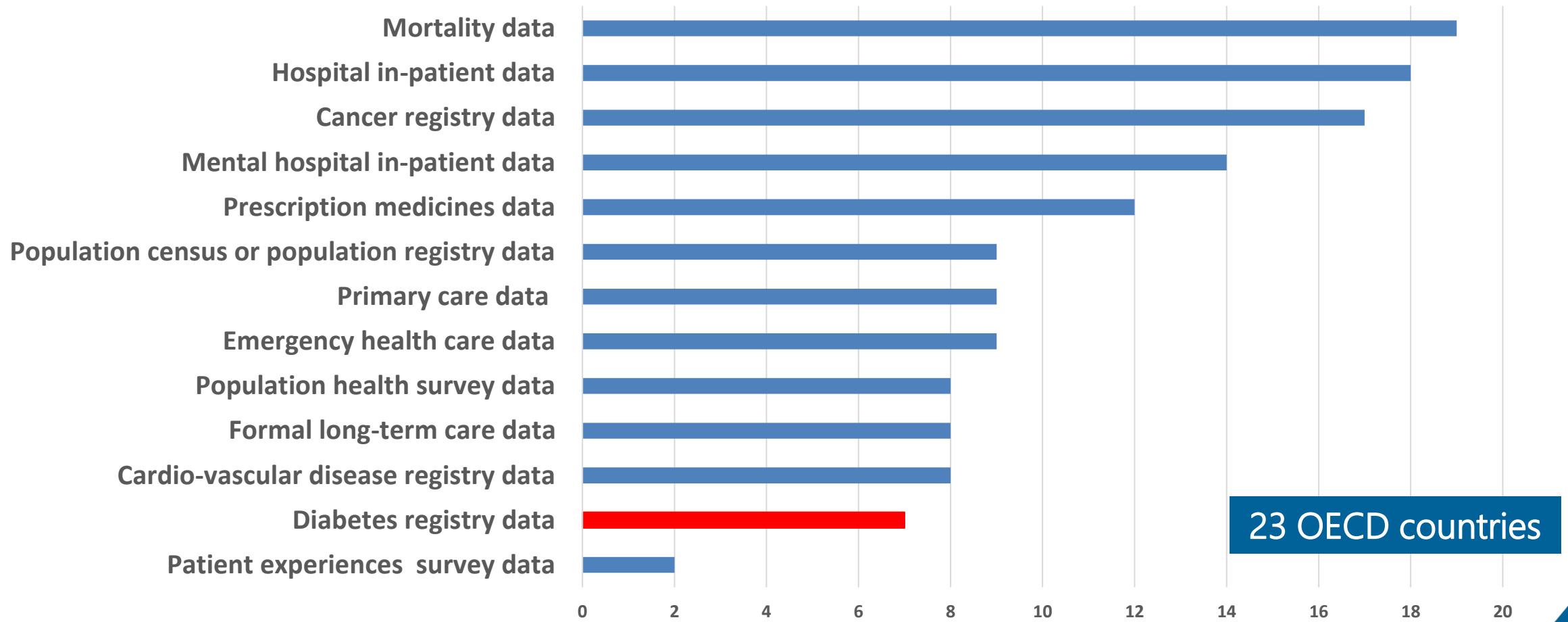


## ...national coverage...



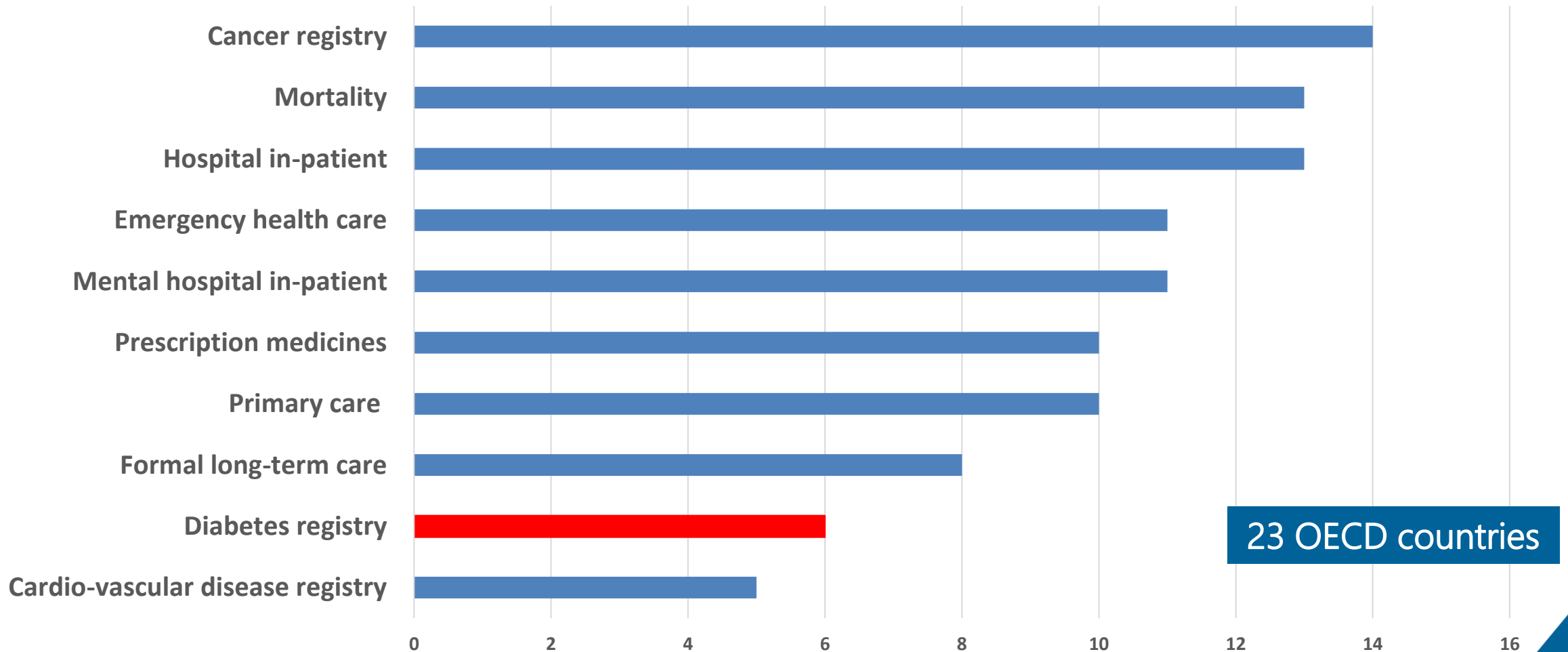


# ...regular record linkage projects...



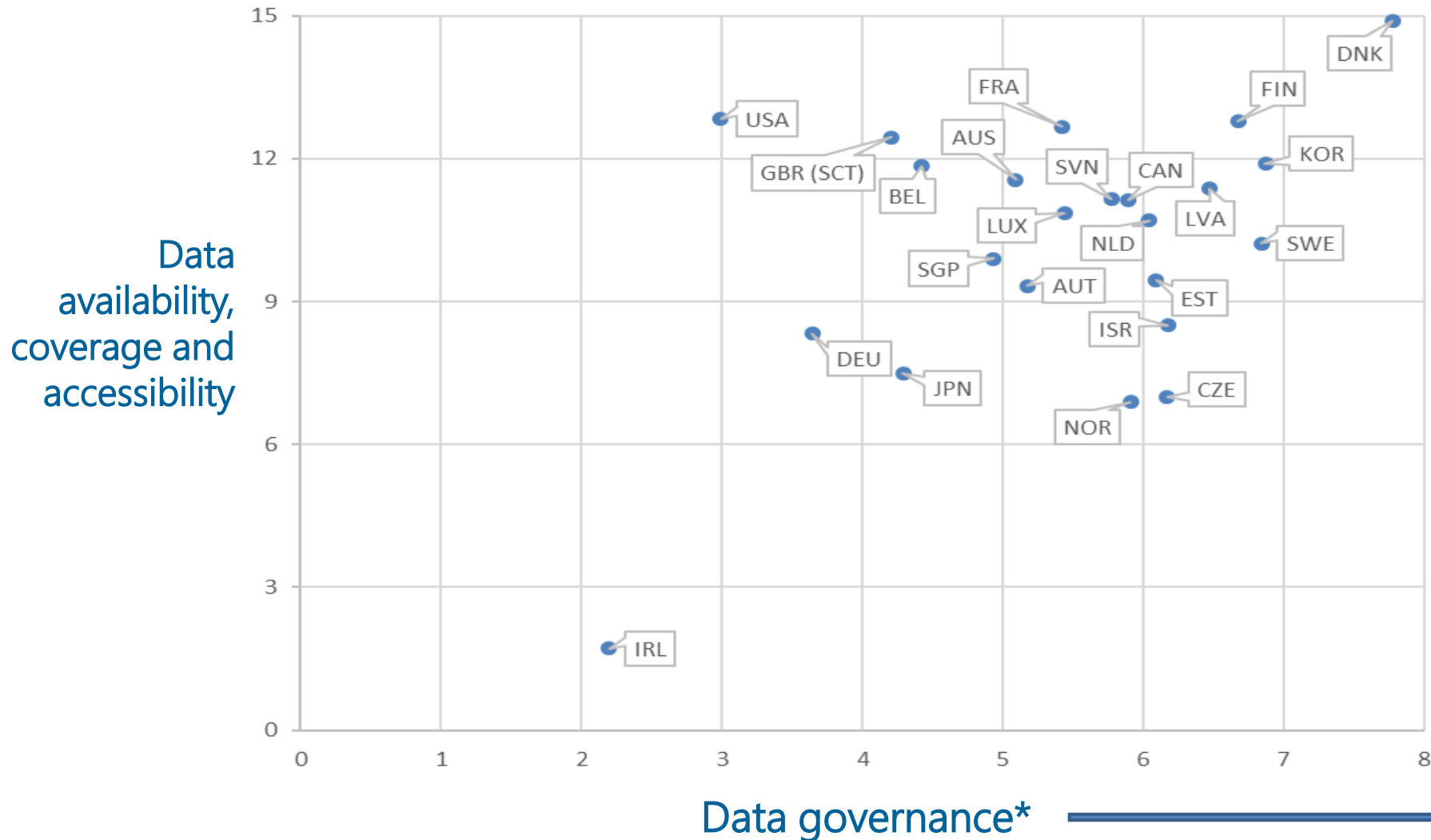


# ... sharing de-identified data with researchers in other countries





# Variation in data use and governance



1. Access with security & privacy safeguards
2. Rules/standards to ensure data quality & interoperability
3. Trust and social license



Roads, rules & regulations, signals,  
proficiency, capability ... trust?



# Things are slowly changing ...

Collaborating on global standards for interoperability



Policies or projects to improve interoperability



Adopting HL7 FHIR standard

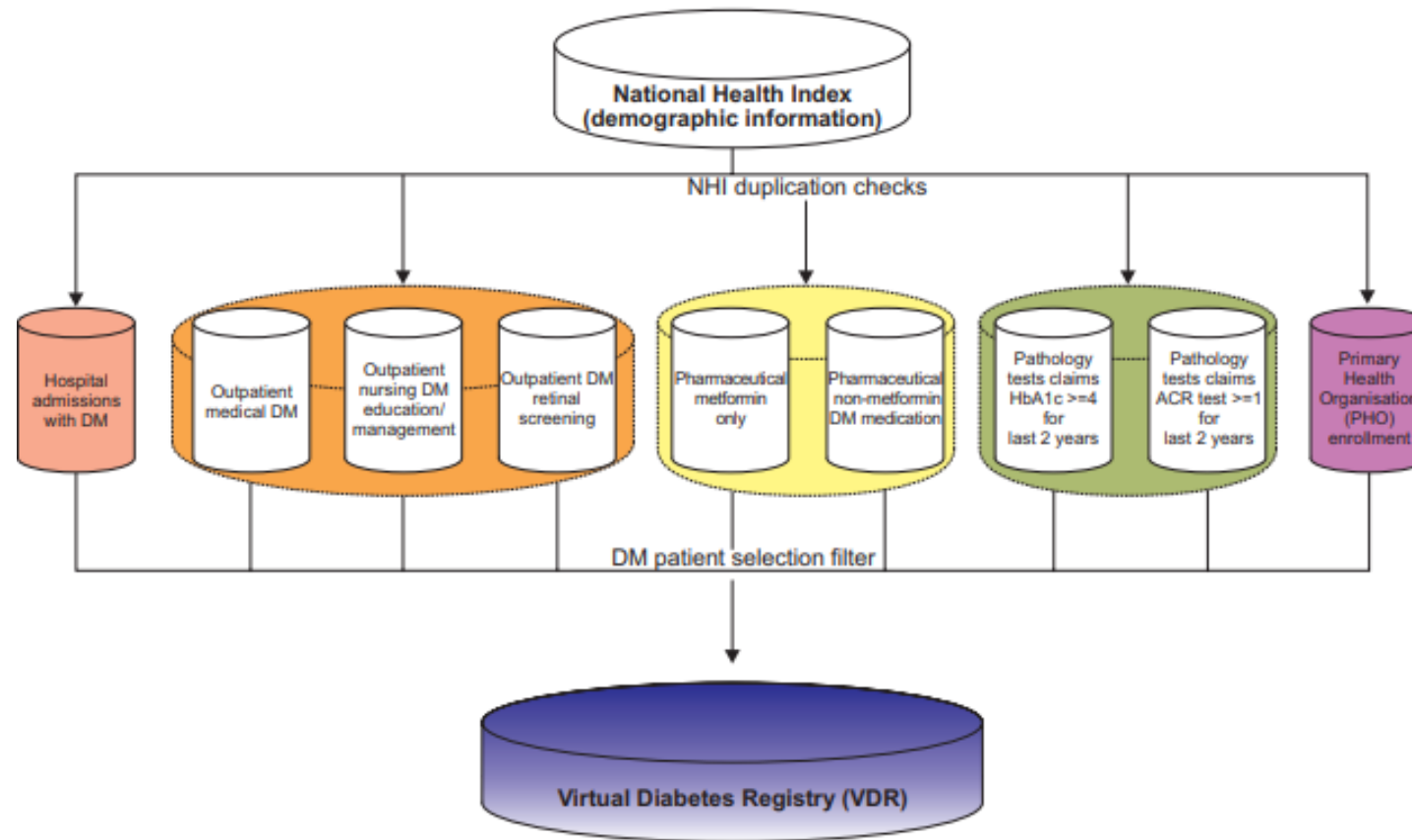


Covid-19 accelerating EHR development and use

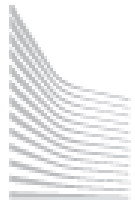




# Virtual diabetes registry (NZ)







European  
Commission

# European health data space

*"...promote better exchange and access to different types of health data (electronic health records, **genomics data**, patient registries etc.), not only to support healthcare delivery ... but also for health research and health policy making purposes."*

## Pillars:

1. Security & privacy (GDPR)
2. Rules/standards to ensure data quality & interoperability
3. Trust and social license

→ i.e. **Governance** (harmonised across MS)



*“to build a common European infrastructure for standardized information exchange in diabetes care, for the purpose of monitoring, updating and disseminating evidence on the application and clinical effectiveness of best practice guidelines on a regular basis”*

Shared European Diabetes Information  
System

SEDIS

<http://www.biro-project.eu/home.htm>



## In summary...

Digital  
technology to  
prevent &  
manage  
NCDs

### - The potential -

Tantalising prospect of 'big data'  
Excellent isolated case examples

### - The reality -

Poor quality data (e.g. bias)  
Held in silos / inaccessible  
Scaling and consistency of tools elusive

### - How to get there -

Health data governance  
quality, interoperable data; securely accessible for 2<sup>o</sup> uses