Bundesinstitut für Impfstoffe und biomedizinische Arzneimittel
Federal Institute for Vaccines and Biomedicines

Das Paul-Ehrlich-Institut ist ein Bundesinstitut im Geschäftsbereich des Bundesministeriums für Gesundheit.

The Paul-Ehrlich-Institut is an Agency of the German Federal Ministry of Health.

www.pei.de

Paul-Ehrlich-Institut

The views expressed in this presentation are not only personal views of the author. They may be understood or quoted as considerations of the Paul-Ehrlich-Institut.

The authors did not receive any funding or financial supplementation, neither by companies nor by Federations representing companies.

Application of Big Data strategies and technologies in proteomics and individualised therapeutic vaccines

Mark Goldammer

21st DGRA
Annual Congress,
23th - 24th May 2019,
Bonn

Application of Big Data strategies and technologies in proteomics and individualised therapeutic vaccines

- Overview
  - Application of Big Data strategies – Chances
  - Application of Big Data strategies – Challenges
- Application of Big Data strategies and technologies in proteomics
  - Relevance of the proteome for medicines development & regulation
  - Proteomics: Methodical challenges – Data Quality
  - Use of Proteomics Data for regulatory purpose – Regulatory Acceptability
- Application of Big Data strategies - individualised therapeutic vaccines
  - Individualised therapeutic vaccines – concepts & examples
  - Requirements – data processing & analysis
- Conclusions

Application of Big Data strategies – Chances

***Volume*[1]**

- Increasing the ***sample size***:
  - *Improving* the tools for ***managing*** (merge) and ***analysis*** of huge data sets, may
    - allow to recognise signals currently not detectable
    - allow to test hypothesis in subgroups currently not available, etc., etc.
  - *Enabling* the collection of huge data sets from ***new sources***
    - e.g. e-health and m-health data
  - Development of new tools for using ***distributed data source*** (*federated data bases*)
    - Often cross boarder transfer of data is not possible due to legal obligations (one of the major problems of ***'data sharing'*** initiatives and projects)
    - The analysis of *federated data bases* is possible ***without physical transfer of data***
    - However, there are currently limitations – new 'smarter' tools for innovative data analysis needed
  - *…*

[1]Volume, Variety, Velocity and Veracity: ***IBM four V's*** – https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Application of Big Data strategies – Chances

Variety

- Integrating of data with a ***variety*** of ***characteristics***
  - New -highly promising- options for data analysis, e.g. ***AI****- or deep learning strategies* allow a (better) analysis of ***multidimensional*** *scientific problems – '****biology is multidimensional'***
    - Research: E.g. improving understanding ***genetic variability*** & impact on (patho-) physiology, pathways or (safety-) pharmacology of medicinal products – *highly complex scientific problems*
  - Analysis of ***'unstructured data'***
    - *'unstructured' -> no CDM: lack of common endpoints, missing data, weakly defined populations,...* BigData strategies hold promises to provide ***'smarter' / more flexible tools for data analysis***
    - Analysis of data with ***limited / unknown quality*** – *'garbage in'* -> '??'
    - ➢ Such approaches / systems need to be validated and to demonstrate ***robustness & validity*** of results
  - *…*

Application of Big Data strategies – Chances

Velocity

- By implementing *automated* ***data processing*** and (*semi-) automated* ***data analysis*** tools *Big Data strategies* allow (in case fully automated systems are established) ***real time analysis*** of data sets.
- The availability of results with substantially reduced or without ***lack time*** will be valuable, e.g.
  - For research and development in general, especially in case a comprehensive and ***timely overview about data from different domains and sources*** is required
  - For pharmacovigilance or drug utility analysis
    - ‘routine’ *pharmacovigilance – signal detection*
    - proactive *safety monitoring, e.g. Sentinel*[1] (US/FDA initiative) – rapid availability of results would be highly valuable and has the potential to improve safety
  - …

Veracity is more a challenge than a chance -> Data Quality

[1]https://www.fda.gov/safety/fdas-sentinel-initiative

Application of Big Data strategies – Challenges

Data Quality

- *Different domains – different requirements and challenges*
  - Clinical (trial) data – best developed domain, requirements well defined (GCP, derogations: Scientific Ad.)
  - RWE – defined requirements for use in pharmacovigilance, other regulatory purposes / licensing (?)
  - Genomics – specific guidance already available in certain fields, e.g. *(companion) diagnostics*
  - (Prote-) Omics – specific guidance required, see below
  - E-health / m-health – regulatory use: under discussion
  - …
- *General recommendations*
  - ➢ implementation of ***quality attributes*** as part or the data set, in order to allow automated identification of adequate data sets for ***data processing*** and analysis
  - ➢ Definition of requirements on data quality, *depending on* ***regulatory purpose***
- …

Application of Big Data strategies – Challenges

Appropriateness of Big Data processing & analysis methods

- ***Which algorithm to use?*** There are many different types of machine learning algorithms - rather than just a single best method.
  The reason lies on the fact that there is no a single method that dominates over all possible data sets and all possible applications. For each particular method there are situations for which it is ***particularly well suited*** *and others where it does not perform as well. These* ***situations are seldom known in advance****, and selecting the best approach can be one of the most challenging parts of performing advanced analytics in practice.*

- Appropriate ***validation*** of Big Data processing & analysis methods is required.

Application of Big Data strategies and technologies in proteomics and individualised therapeutic vaccines

- Overview
  - Application of Big Data strategies – Chances
  - Application of Big Data strategies – Challenges
- Application of Big Data strategies and technologies in proteomics
  - Relevance of the proteome for medicines development & regulation
  - Proteomics: Methodical challenges – Data Quality
  - Use of Proteomics Data for regulatory purpose – Regulatory Acceptability
- Application of Big Data strategies - individualised therapeutic vaccines
  - Individualised therapeutic vaccines – concepts & examples
  - Requirements – data processing & analysis
- Conclusions

Relevance of the proteome for medicines development & regulation

The genome is static – the *proteome is* ***dynamic***

The proteome is relevant for

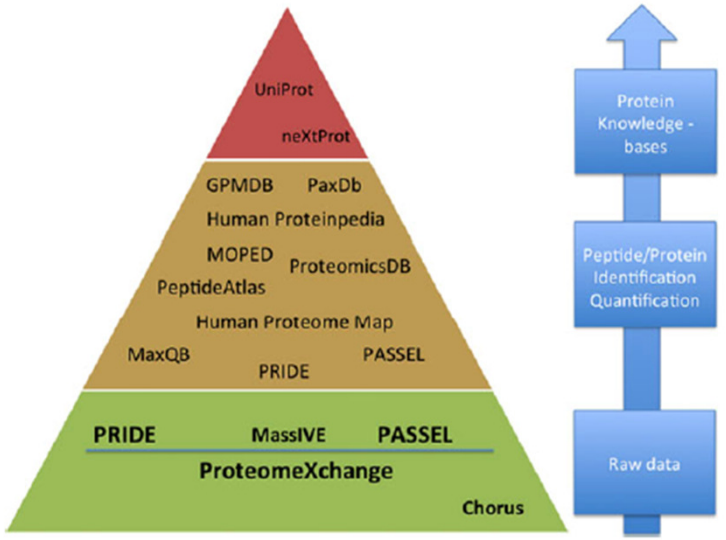
- *Physiology & Pathophysiology*

Proteins are ***targets*** for most medicinal products & are relevant for

- *Pharmacology – efficacy*
- *On-target safety pharmacology (immuno-tox)*
- *Off-target toxicology*

➢ High-throughput proteomics holds substantial promises to improve medicines development & regulation

Changes of the proteome ***do not necessarily*** correlate with variability in the ***genome***…



proteomics data repositories and databases according to the different data types stored

Proteomics[1]: Methodical challenges – ***Data Quality***

The *proteome is* ***dynamic and multidimensional***

The *proteome is* ***highly variable*** – inter-individually, but also intra-individually: depending on the *the sample source*, *time point of sampling*, *environmental and behavioral factors*, *medical conditions* and *treatments*, etc.

The *concentration* of individual proteins *in complex matrices is* ***highly variable***:

- Measuring proteins with ***low concentrations*** *maybe very challenging*
- The lower limit of quantitation (***LLOQ***) is a very important factor – important ***quality attribute***
- Especially the analysis of ***complex matrices*** *is highly challenging* (variability of the matrix)

Ca. 30% of a 'typical' proteome are ***membrane proteins*** – the quantification of membrane proteins is associated with specific methodical challenges, like ***aggregation*** *or* ***precipitation***

Serum Proteomics: Serum is one of the ***most complex proteomes*** *in a complex, highly variable matrix*

[1]***parallel analysis*** of huge numbers of different individual proteins

Use of Proteomics Data for regulatory purpose – Regulatory Acceptability

Quality and variability of proteomics data

- Method *validation* and demonstration of (required level of) *validity* of data

Reproducibility *(robustness – matrix effects/variability)*

- There is a general need for methodical improvements, with a *focus on reproducibility*

Data standards

- In order to enable automated data processing and analysis there is a need to include ***quality attributes*** in the meta data of relevant data sets – including the necessary information for the regulatory use of the data sets.

Regulatory guidance

- *Guidelines*, defining the requirements on data quality – depending on intended *regulatory purpose*
- However, there will be the need for *more specific guidance* for particular product & method
  - *Scientific Advice*
  - *Qualification Advice*

Application of Big Data strategies and technologies in proteomics and individualised therapeutic vaccines

- Overview
  - Application of Big Data strategies – Chances
  - Application of Big Data strategies – Challenges
- Application of Big Data strategies and technologies in proteomics
  - Relevance of the proteome for medicines development & regulation
  - Proteomics: Methodical challenges – Data Quality
  - Use of Proteomics Data for regulatory purpose – Regulatory Acceptability
- Application of Big Data strategies - individualised therapeutic vaccines
  - Individualised therapeutic vaccines – concepts & examples
  - Requirements – data processing & analysis
- Conclusions

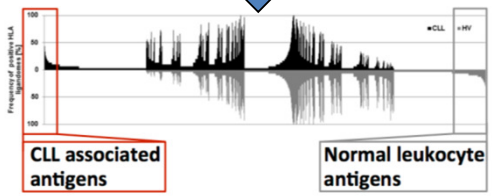
Individualised therapeutic vaccines – concepts & examples

by **LC-MS/MS**

Proteomics

Bioinformatic analysis

HLA ligandome analysis tumor tissue



Mutanome sequencing tumor tissue

- Prediction of ***patient-individual neo-expitopes***
- Neo-Epitope selection and peptide design for manufacturing

by **'Epitope Prediction'**

Genomics

Bioinformatic Algorithms

Peptide Warehouse

Mutated Peptides

Selection of peptides (***medicinal products***) relies on actual ***presentation of HLA ligands*** in patient tumour samples

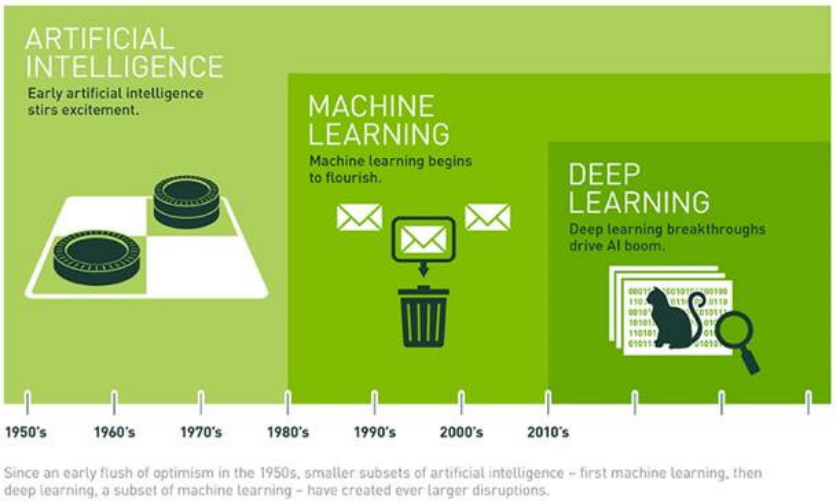
off the shelf

GAPVAC trial

***de novo* synthesis**

Selection of individualised ***medicinal products*** entirely relies on **bioinformatic algorithms**

Requirements – data processing & analysis



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

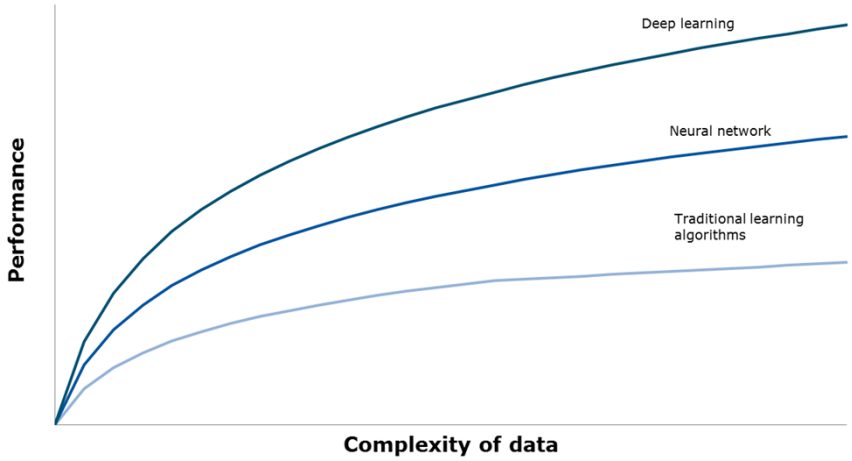
NVIDIA, https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Traditional programming	Machine learning
Pre-programmed by humans: producing the same results every time	*Machine learning*: changing its code based on results
Deterministic: choices are clearly defined	*Stochastic*: based on probability
One-dimensional: for one/limited purpose	*Multi-dimensional*: potential for more general purposes

Machine Learning Security, https://www.slideshare.net/eburon/machine-learning-security-ibm-seoul-compressed-version

Nature / Biology is *'multi-dimensional'*

Requirements – data processing & analysis



Andrew NG Deep learning course. The graph is conceptual and serves as illustration, presentation by Rigg J. Using machine learning and real world data to tackle complex healthcare challenges. IQVIA April 2018

Choosing the right model:

- There are now situations where the quantity and ***complexity of data*** available is ***beyond*** not only the ***cognitive capabilities of the human*** brain ***but also*** of the more ***'traditional' analysis methods***…and access to this high dimensional data is critical to enable better predictive accuracy.
- Different to deterministic machine learning algorithms, ***non-deterministic*** approaches like deep learning ***do not provide the same answer if they are re-exposed to the same data set***.
  It is difficult to *differentiate* (and to predict if this will be also the case in the future) – if this may *indicate limited reproducibility* of an approach, or it is *just a correct (better) answer*.
- …
- Even if there is an attraction, it ***is wrong*** to assume that always the ***most sophisticated*** models / machine learning techniques should be used
  "make your model as simple as possible, but not simpler" – Albert Einstein

Data Analytics – Analytical methodologies. *Subgroup report, Joined HMA/EMA – BigData Taskforce*

Requirements – data processing & analysis

Some particular characteristics of Big Data processing & analysis methods

- ***Over-fitting**: Some of the machine learning models available are **very flexible** and provide **nearly perfect** results on the **data they have been trained** on, but, **less accurate predictions on new observations** (lack of external validity - generalisability of the model).*
  - *Techniques to avoid over-fitting are called regularisation techniques - an area of intensive research.*

- Today, most AI analysis approaches are optimised to ***'work backward from data to patterns or relationships' (data fishing)[1]** – they are generally not design to confirm results:*
  - This qualifies these methods for ***signal detection*** – a very useful approach in pharmacovigilance
    - Are there other regulatory applications these properties would be useful?
  - Can ***AI analysis approaches*** be developed that ***fulfil requirements to use them for confirmatory regulatory purpose*** (licensing)?
    - It is necessary to define this requirements – *Regulatory Acceptability*

[1]Clare Gollnick, CTO of Terbium Labs – https://towardsdatascience.com/the-reproducibility-crisis-and-why-its-bad-for-ai-c8179b0f5d38

Requirements – data processing & analysis

Validation of Big Data processing & analysis methods

Challenges

- It is challenging to ***document*** the steps through each iteration of deep learning approaches
- Data science stack because ***deep learning has a lot of 'moving parts'***,
  - and changes in any of the different layers of the deep learning framework
- Characteristics of ***GPU*** (Graphics Processing Unit) an there driver software have to be taken into account *(the mathematical basis of neural networks and image manipulation are similar)*
- ***Training or validation datasets*** – can all impact results
- …

Challenges for regulator

- Defining requirements, depending on regulatory purpose – ***Regulatory Acceptability***

Application of Big Data strategies and technologies in proteomics and individualised therapeutic vaccines

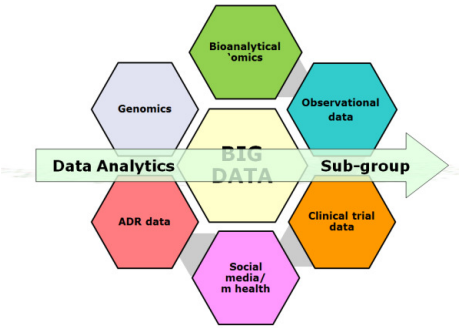
- Overview
  - Application of Big Data strategies – Chances
  - Application of Big Data strategies – Challenges
- Application of Big Data strategies and technologies in proteomics
  - Relevance of the proteome for medicines development & regulation
  - Proteomics: Methodical challenges – Data Quality
  - Use of Proteomics Data for regulatory purpose – Regulatory Acceptability
- Application of Big Data strategies - individualised therapeutic vaccines
  - Individualised therapeutic vaccines – concepts & examples
  - Requirements – data processing & analysis
- Conclusions

Conclusions

Application of Big Data strategies and technologies holds substantial promises for research and ***development of innovative medicines*** – but also ***therapy optimisation***.

Chances and challenges have to be evaluated in order to develop a ***suitable and innovation friendly regulatory framework****:*

- ✓ *Enabling the use of promising technologies*
- ✓ *On the same time ensuring safety and benefit for patients and society*



Joined HMA/EMA BigData Taskforce [stage I]

There should be a specific focus on:

- Availability of ***adequate data*** sets – depending on the intended regulatory purpose
- Tools for ***data processing and analysis*** – ***sufficiently validated*** & fit for purpose

➢ In order to allow drawing conclusions which are ***regulatory acceptable***


Questions?