Planning Future Health: Developing Big Data and System Modelling Pipelines for Health System Research

Niki Popper[1*], Florian Endel[2], Rudolf Mayer[3], Martin Bicher[2], Barbara Glock[1]

[1]COCOS – Centre for Computational Complex Systems, TU Wien, Wiedner Haupstrasse 8-10,
1040 Vienna, Austria; **nikolas.popper@tuwien.ac.at*

[2]DEXHELPP, Neustiftgasse 57-59, 1070 Vienna, Austria; [3]SBA Research, Favoritenstrasse 16, 1040 Vienna, Austria

SNE 27(4), 2017, 203 - 208, DOI: 10.11128/sne.27.tn.10396
Received: December 2, 2017; Revised: December 15, 2017;
Accepted: December 22, 2017
SNE - Simulation Notes Europe, ARGESIM Publisher Vienna,
ISSN Print 2305-9974, Online 2306-0271, www.sne-journal.org

Abstract. Using Modelling and Simulation for Planning and Evaluation of Health Systems needs the combination of state of the art modelling and simulation tools, which can be used in a modular way. In addition, big data sets are used to describe the status-quo of health systems. Together possibilities for prediction and strategic planning increase, while challenges for quality assessment and reproducible processes generate the need for new developed structures in research and implementation. The article gives an outline of approaches, developed within the Austrian DEXHELPP project and points out strategies for coping with these challenges.

Introduction

Based on the concepts of equations, networks, algorithms and causal understanding of the world, modelling and simulation has reached a high level of describing systems and processes, like complex technical systems, ecological systems or production and logistic processes. Challenges are developing, combining, transferring and communicating such approaches and solutions.

On the other hand, big data has come to an eminent importance based on sensors and computational efforts in measuring our world. Today, many technologies for the construction, monitoring, and evaluation exist. A huge amount of activities can be seen in research, founding, policy and media.

Still, interfaces and methods for linking these technologies are to be intensified. Especially complex socio-technical systems intertwining technologies and humans, aiming to serve the goals of citizens on different levels, ranging from the personal goals of a citizen regarding, e.g., treatment and prescription management to the management of Health Care by policy and decision makers. One prerequisite to achieve this is the analysis of actual data and forecasting of the future behaviour with simulation methods.

1 Future Health Systems

National health systems invest annually more billions of Euros as the demand for health services increases (because of demographic change), but resources are limited. In addition, complexity of processes increases: From diagnosis to therapy – Interventions are complex hybrid processes including e-health, decentralised services and personalised medicine.

Measurement of efficiency and effectiveness becomes more and more complex but is an urgent need. Development of new methods, models and technologies is needed to support analysing, planning and controlling. Quantity and quality of available data strongly increases and therefore facilitates the description and analysis of all areas in complex systems like health care. Based on data for healthy expenditures the evaluation for health care systems has a market of 75 to 120 Billion Euro only in the European Union.

To provide state of the art analysis for Health Technology Assessment (HTA), Comparative Effectiveness Research (CER) and Evidence Based Medicine (EBM) processes combining health system domain knowledge, knowledge of professional data processes and – finally – mathematical modelling & simulation will be vital to transform Big Data into Deep Data: Evidence based and reproducible knowledge (See [1,2]).

Bringing together these technologies is an enormous challenge. Data Based Demographic models must be combined with models for the spread of diseases. Time dependent treatment paths must be parametrized with data sets from clinical routine joined with large scale health system data. For system simulation an important aspect is the possibility to implement changes inside the system, like interventions within the computer model, and to analyse their effects.

On basis of experiences of the Austrian DEXHELPP COMET project for Decision Support in Health Policy and Planning, where an innovative research infrastructure was developed to enable researchers and other stakeholders to share data and methods for research and decision making, some important points to handle these complex processes in future will be described.

2 Big Data and Complexity

Next generation tools are needed to make development, construction monitoring, and analysis of such systems easier, faster, more reliable and – one of the most important things – comprehensible for decision makers and other stakeholders. To achieve these goals, methods in the following areas must be developed:

- **Data:** Integration, storage, management and analysis of very large amounts of data, unstructured data, secure and reproducible data management from sensors, IoT and various data sources like dynamic databases or non-structured information sources, collecting data, interfaces and analysis methods from statistics, machine learning, and visual analysis.
- **Models:** Formal, scalable modelling of various systems, heterogeneous modelling of subsystems and the integration of these subsystems, development of modelling methods for computational complex systems, multi method modelling, including coupling and comparison based on data, system knowledge and applications demands.
- **Computation:** Combining data and models to simulation tools for complex systems, developing innovative methods in numerical mathematics, co-simulation, hybrid simulation and reproducible simulation.
- **Interfaces:** Interfaces and visualization of simulation results, decision support systems and future development of Human-Computer Interaction (HCI).

3 Data Infrastructures

Data required in scientific investigations in the health domain is in many cases difficult to share, as it contains personalized information, e.g. on patients and service providers in the health sector. However, detailed and high-grade data is required to compute high-quality mathematical and statistical models and visualizations of interesting facts.

To facilitate sharing of data, infrastructure, software tools and well-defined processes must be developed to enable researchers and customers to share their data in a secure and transparent way. DEXHELPP developed a prototype for support of Austrian health system research, sharing information between research community and the health system stakeholders.

3.1 Infrastructure for research

Computational sciences such as e-Health depend heavily on the detail and quality of the data that is utilised to develop mathematical and statistical models. However, especially with data in the health sector that often comprises sensitive data on people, protection of the privacy of these individuals is often impeding easy exchange of data between data holders and consumers.

DEXHELPP tackles this problem by having created a secure and controlled environment (see below Figure 1) where data holders can deposit their data, and data consumers can perform their analysis and experiments within that environment, without the need to transfer the data outside of the system.

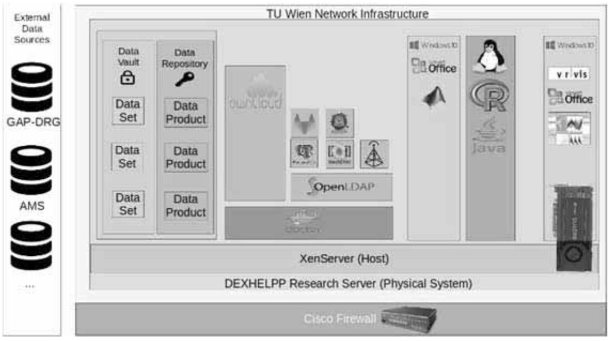

Figure 1: Overview of the architecture of the DEXHELPP Research Server infrastructure.

Data providers can specify fine-grained access rights to individuals or groups of researchers, to complete data sets or just specific subsets thereof, e.g. limiting the number of records, or excluding specific details of records.

The access of data consumers to these sources is accurately recorded, which allows for auditing and inspection of the intended usage of the data.

One important aspect of the system is the trade-off between the controlled environment, and the choice and offer of modelling and programming tools available to the researchers. We tackle this by providing the researchers a wealth of commonly used tools, the requirements for which were elicited by observing current practices. Further, the server environment offers a fast computing environment, with special hardware such as GPU computing available on demand.

3.2 Active and passive security measures

Active security measures aim at preventing access to data and other resources by unauthorised parties. In the DEXHELPP infrastructure, these include:

- Basic security provided by the Vienna University of Technology network
- A central authentication and authorisation schema, based on a directory service.
- A dedicated Virtual Private Network (VPN), which provides full encryption of all network transfer between the users and the infrastructure. The access to the VPN is protected with a two-factor authentication.
- Provisioning of access to data and resource only in the granularity required for the specific research question.

Passive security measures enable detection of inappropriate access and usage of DEXHELPP resources. They include

- A centralised and replicated logging of access to all resources, including services and data used by DEXHELPP partners and DEXHELPP customers.
- Embedding watermarks and personalised fingerprints into the data allows detecting data leakages, and to attribute the leakage to a specific user account (Figure2).

3.3 Impact and effects

With the DEXHELPP research services infrastructure, both data providers and data consumers profit from the facilitation in the exchange of and access to data, which dramatically speeds up the bootstrap process of a new research investigation. Due to the integrated data sharing and analysis infrastructure, data providers have the assurance that access to the data is accurately recorded and can be tracked and inspected also retroactively.

No. Bl. Type	Rhes.	Diagnosis Prescript	Link Type

Figure 2: Active and passive security measures, like fingerprinting of data sets, are combined to unite privacy issues with possibilities for research and policy.

The usage of the secure research server infrastructure increases the trust among the project partners and increases the readiness for sharing data [3].

Data consumers benefit from an easier access to the data, the powerful computing environment, and, with security measures provided centrally as a service, can focus on their skills and actual tasks at hand, instead of spending time on data privacy and security concerns. The central provisioning of data and services also facilities the usage of tools that enhance the reproducibility of research investigations.

The research server is also a suitable environment for a third party to merge and link data sets from different sources, which otherwise would not be released, for example combination of data on social insurance with data on the labour market.

4 Population Models

Simulation models for aspects of the health system often require an underlying population model, which require substantial resources in the modelling process. In DEXHELPP a '**Ge**neric **Po**pulation **C**oncept' (GEPOC) was developed to collect population data and implement population models to have them readily available.

Models are designed to be extended for healthcare aspects to make future projects faster and more efficient. Above that, this allows reproducible use of data and standardized transferability to other countries. As proof-of-concept, GEPOC was applied on various applications, like a vaccination model.

4.1 Motivation

Simulation models are a common way to perform analysis and planning in the healthcare system.

Often, data does not directly provide answers and statistical analyses are insufficient. In the past DEXHELPP partners conducted several projects in Austria that included modelling and simulation. Agent-based models were developed to simulate the outcome of vaccinations for streptococcus pneumoniae and influenza.

Another model analysed cost effectiveness of screening for Abdominal Aortic Aneurysms. In another project, an agent-based model was developed to compare the costs of healthcare in different reimbursement systems. These models addressed completely different issues and answered completely different questions. However, each model needed a valid representation of Austria's population, and structures to simulate changes like births or deaths. Each project spent substantial resources to do the same thing again and again.

Therefore, the GEneric POpulation Concept – GEPOC – should provide appropriate structures and model parts about the Austrian population. This should substantially decrease the effort of population modelling, and further the total effort of modelling projects in the future.

4.2 Outline

GEPOC generally consists of three parts: A handbook, model implementations, and data. The handbook contains a literature review on other population models, lessons learned from past projects, and comprehensive documentation of collected data and implemented models.

Model structures were implemented as agent-based models in Java and Python, and as a System Dynamics (SD) model in Anylogic. These combinations of implementations serve a wide range of projects conducted by DEXHELPP partners and customers. All models can simulate a population of humans with age and sex, including births, deaths, immigration, and emigration over a period of several years.

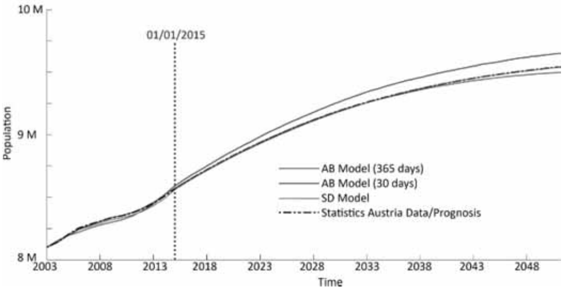

Figure 3: The resulting population of the AB model (with 365 and 30 days time steps) and the SD model compared to Statistics Austria's data.

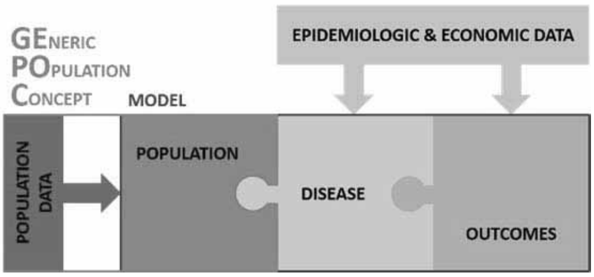

Figure 4: Models based on GEPOC can re-use population data to enrich these data sets with additional data.

The data about Austria's population, including births, deaths, immigration, and emigration, was collected from Statistics Austria. The data composes of data until 2014 and predictions until 2050. The total population's prediction is a result of Statistics Austria's own population model. The handbook describes the sources of this data, the pre-processing and storage format, and the mathematical calculations for model parameters.

4.3 Validation

Both the AB and the SD model were simulated for Austria's population from 2003 until 2050, using the initial population in 2003 and births, deaths, immigration, and emigration data for 2003-2050.

For validation, the resulting model population was compared to the total population data. Figure 3 shows that this works well. The SD model exactly resembles Statistics Austria's population prediction. The AB model contains small differences, which are acceptable and trace back to complex parameter calculation. Further research addresses this issue to minimize the differences.

4.4 Impact

The population models can reproduce Statistics Austria's data. Thus, they can be considered valid and serve as a basis for further modelling. This is the function that the pure population data cannot provide. The models are designed in such a way that they can easily be extended for diseases, treatments, and other healthcare interventions (Figure 4).

The population model, which serves as a necessary core of the model, can perform a standalone simulation of the population. Additional models were implemented [4, 5], e.g. an agent-based Python model to simulate vaccination rates in Austria for the Austrian Ministry of Health.

Information was available about yearly vaccination numbers in Austria, disease outbreaks, as well as vaccination state of immigrants and refugees. As a result, the model shows vaccination and immunity states of Austria's population for each birth cohort. This analysis allows performing specific measures for children vaccinations and catch-up vaccinations to reach the health goals concerning the disease.

GEPOC provides population data readily available and allows using it in a well-defined reproducible way. The implemented models are designed to be used for healthcare issues and significantly reduce resources, time and cost of modelling projects. Data and models can easily be updated whenever new data is published. Data handling and models can even be transferred to simulate other countries without any major changes. Using the DEXHELPP Research Server and GEPOC provides so the possibility to implement fast and reproducible health system models.

5 Atlas of Epidemiology

In healthcare it is crucial to have a fundamental knowledge of the burden of diseases within the population. Therefore, one aim is to develop various Atlases, like an Atlas of Epidemiology to gain better insight on the epidemiological situation. Based on primary and secondary health care data, the goal was to present results in interactive charts and maps, comprehensible to experts and the public. The atlas builds a framework for rapid deployment of new data and results in a reproducible and efficient way. As first use case three methods based on two different databases for the estimation of diabetes prevalence in Austria were compared.

5.1 Different sources – different results: methods applicable for the public

Evaluating the burden of disease by estimating the prevalence or incidence of a disease within a population is usually conducted based on a specific data source. Often different data sources exist and lead to different – sometimes horrendous different – results, which makes decision making and further analyses difficult. Communication the different results to the public and explaining these differences seem to be challenging. Therefore, the partners of DEXHELPP developed concepts and a first use case of the Atlas of Epidemiology, a web-based tool, capable of tackling this challenge.

The tool developed here is generic and can be adapted to other questions of epidemiology of interest. The first use case compares three methods of estimating the prevalence of diabetes based on two different data sources. The success of this use case and the positive response by social security institutions and individuals, encouraged us to plan and (currently) develop a second use case: the prevalence of different diseases according to the ICD10 chapters based on the ATC-ICD project, where diagnoses are assigned to patients based on hospitalizations, sick leaves and prescription data.

The methods used for communicating the different results of the burden of disease to the public include interactive bar charts and heat maps of Austria for patients divided by region, gender and age together with descriptions of results. The user is encouraged to experiment with the different results.

5.2 Successful implementation: use case Diabetes

The first use case for estimating prevalence of diabetes was successfully implemented based on two different data sources:

I. reimbursement data 2006/2007 (GAP-DRG);
II. national routine health surveys (ATHIS) for 2006/2007;

with three different methods (see Figure 5):

1. ATC-ICD statistically relates pseudonymized data on medications to data on diagnoses from hospitalizations and sick leaves.
2. Experts: medical experts assign specific medications to diabetes diagnoses. Patients with these medications are identified together with hospitalized diabetes diagnosed patients in GAP-DRG.
3. ATHIS a sample of 15.000 persons was questioned if they a) ever had diabetes and b) were treated against diabetes in the last 12 months. Results are projected onto the Austrian population.
4.

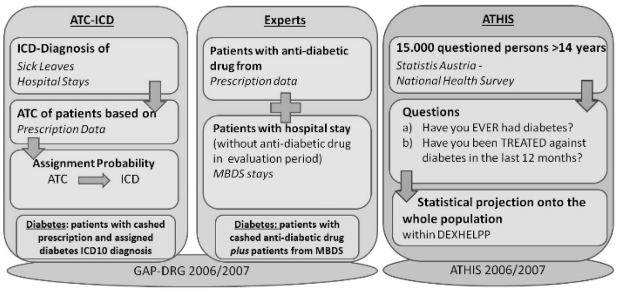
ATC-ICD
ICD-Diagnosis of
Sick Leaves
Hospital Stays
ATC of patients based on
Prescription Data
Assignment Probability
ATC ICD
Diabetes: patients with cashed prescription and assigned diabetes ICD10 diagnosis
Experts
Patients with anti-diabetic drug from
Prescription data
Patients with hospital stay
(without anti-diabetic drug in evaluation period)
MBDS stays
Diabetes: patients with cashed anti-diabetic drug plus patients from MBDS
GAP-DRG 2006/2007
ATHIS
15.000 questioned persons >14 years
Statistis Austria -
National Health Survey
Questions
a) Have you EVER had diabetes?
b) Have you been TREATED against diabetes in the last 12 months?
Statistical projection onto the whole population
within DEXHELPP
ATHIS 2006/2007

Figure 5: Different Methods and Different Data Sources for the Use Case of Diabetes Prevalence.

Patients are divided by 10-year age-classes, gender and region. For the publicly online framework (http://www.dexhelpp.at/de/epidemiologie-atlas/), implemented in html and javascript, pre-processed data in different granularity is required and used.

5.3 Impact and effects

The atlas was publicly presented online to decision makers, health care providers and social security systems [6]. The feedback was very positive to our novel approach and has been incorporated. Maps of Austria represent the prevalence of diabetes for each method and granularity level. The difference of the methods can be seen by clicking on the next map. For different age-classes (resp. different gender) the three methods can be compared directly within a bar chart. The technology for a rapid deployment of new data is now developed which has a major impact on estimating the burden of disease, since it is now possible in a fast an efficient way which allows focusing more on the intervention than the representation of epidemiology.

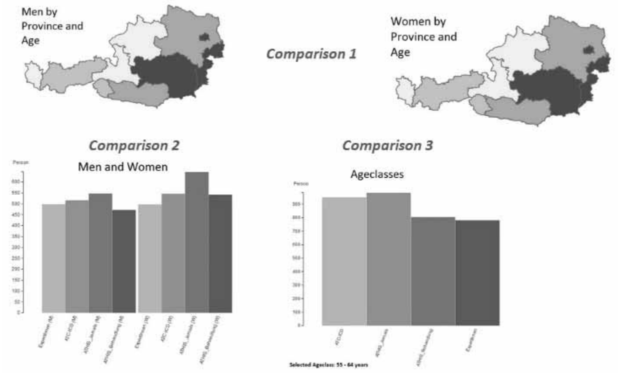

Figure 6: Interactive Comparison methods for Use Case of Estimating Diabetes Prevalence based on different data sources.

Besides depicting disease prevalence, the Atlas of Epidemiology also allows to quickly visualize health care service data and results of simula-tion models, which is important for decision makers and the communication to the public. Results of the ATC-ICD project on the prevalence of different diseases based on ICD9 diagnoses and medication data were also integrated in an aggregated form.

6 Big Data and System Modelling

Three examples for Data Processes, Modelling & Computation and possible Interfaces were described above.

These are only examples for methods already implemented in DEXHELPP and many other activities in Europe and around the world. The near future will bring integrated solutions with Deep Learning, Network Models and many more. E.g. based on modules like GEPOC and the DEXHELPP Research Server, virtual clinical trials can be implemented soon.

References

[1] Habl C, Renner A, Bobek J, Laschkolnig A. Study on Big Data in Public Health, Telemedicine and Healthcare. Report / Study, 16.12.2016, Directorate-General for Health and Food Safety, Directorate B—Health systems, medical products and innovation.

[2] Salcher M. Connecting the Dots: Putting Big Data to work for Health Systems. Eurohealth; Quarterly of the European Observatory on Health Systems and Policies Vol.23, No.1 2017; S.3-6

[3] Pröll S, Rauber A. Enabling Reproducibility for Small and Large-Scale Research Data Sets. D-Lib Magazine, 23 (2017), 1/2; 6 S.

[4] Bicher M, Urach C, Zauner G, Rippinger C, Popper N. Calibration of a Stochastic Agent-Based Model for Re-Hospitalization Numbers of Psychiatric Patients. Proc. 2017 Winter Simulation Conference, Chan W K V, D´Ambrogio A, Zacharewicz G, Mustafee N, Page S (eds.); IEEE, CFP17WSC-USB (2017), ISBN: 978-1-5386-3428-8; 12 S.

[5] Schneckenreither G, Popper N. Dynamic Multiplex Social Network Models on Multiple Time Scales for Simulating Contact Formation and Patterns in Epidemic Spread. Proc. 2017 Winter Simulation Conference, Chan W K V, D´Ambrogio A, Zacharewicz G, Mustafee N, Page S (eds.); IEEE, CFP17WSC-USB (2017), ISBN: 978-1-5386-3428-8; 12 S.

[6] Glock B, Endel F, Endel G, Sandholzer K, Popper N, Rinner C, Duftschmid G, Filzmoser P, Mert M C, Holl J, Wagner-Pinter M. How sick is Austria? - A decision support framework for different evaluations of the burden of disease within the Austrian population based on different data sources. Int. Journal of Population Data Science, 1 (2017), 1; S. 92.

Note. The contribution was supported by DEXHELPP Decision Support for Health Policy and Planning, which is funded by the Austrian Ministries for Innovation, Technology, Science, Research and Economy as well as the state of Vienna via COMET (Competence Centers for Excellent Technologies). The COMET program is processed by FFG (Austrian Research Promotion Agency).