Deliverable 2.3

Tools for estimation of relevant parameters

Project acronym: *DYNANETS*

Project full title: Computing Real-World Phenomena with Dynamically Changing Complex Networks.

Grant agreement no.: 233847

Due-Date:	24
Delivery:	24
Lead Partner:	EMC
Dissemination Level:	Public
Status:	Final
Approved:	Q Board, Project Steering Group
Version:	9.0

DOCUMENT INFO

0.1 Authors

Date and version number	Author	Details
April 18th 2011 v1.0	David van de Vijver (EMC)	Outline of all the work done. Shared with partners.
April, 19th 2011 v2.0	David van de Vijver, Dineke Frentz, Imke Schreuder, Brooke Nichols (EMC)	Introduction and case studies included
April, 26th 2011 v3.0	Mattia Prosperi (UCSC)	Contribution of UCSC
April, 27th 2011 v4.0	Rick Quax (UvA)	Contribution of UvA, with description model and Amsterdam cohort study
May, 20th, 2011 v5.0	Lorenzo Isella (ISI)	Contribution of ISI, contact networks
May 20th, 2011 v6.0	K. Soja, J. A. Hołyst, J. Sienkiewicz (WUT)	Work on university ranking
May 27th, 2011 v7.0	L. Isella (ISI), R. Quax (UvA), D. van de Vijver (EMC)	Part on handling missing data
May 30th, 2011 v8.0	David van de Vijver (EMC)	Executive summary and discussion
May 31st, 2011 v9.0	Janusz Holyst	Comments on the summary

TABLE OF CONTENTS

List of Figures and Tables

1 Executive Summary

Objectives

The objective of this deliverable is to analyze datasets available to DynaNets in order to uncover correlations between the structure of the network on the one hand and the nodes features on the other hand. In addition, we report on university rankings. Finally, we report on techniques for handling missing data.

Context of this deliverable within workpackage 2

This deliverable has three different parts. First, we used datasets described in deliverable 2.1. These datasets were analyzed using statistical-, mathematical, and phylogenetic-techniques. The latter includes sophisticated tools that were reported as part of deliverable 2.2. The results of the analysis that are reported in the first part will be used in deliverable 2.4. We chose three case studies with practical epidemiological and clinical objectives. (These case studies allowed us to convince data providers to collaborate with us). In the second part, we report on the results of the university rankings. In the third part, we report on several methods that can be used for handling missing data.

Part 1 – analysis of datasets

Case studies

The case studies that we chose are 1) the impact of super spreaders on the HIV epidemic, 2) impact of earlier treatment on the spread of HIV and 3) future levels of transmitted drug resistance. We will study these objectives using an agent-based computational model for simulation of the spread of HIV. The agents are connected with a dynamic complex network in which each agent represents an individual MSM (men-having-sex-with-men) with individual infection status (healthy, acute stage, chronic, AIDS, treated) and behavior (sexual partnerships).

Results

Below is a summary of the main results. These results will be incorporated in the model

Analysis	Main result
Biology of HIV infection	Summary of literature concerning transmissibility and disease progression.
Phylogenetic transmission networks	Transmission of drug resistance most frequently due to onward transmission between antiretroviral naïve patients and forward transmission from patients failing treatment to newly infected patients is

Analysis	Main result
	uncommon.
Clinical care	A substantial part of HIV infected patients are only diagnosed at an advanced stage of their infection. These patients that are unaware of their infection may continue to spread (drug resistant) HIV to others.
Virological failure	Prediction of time to virological failure.
Transmission of drug resistance	Stable over-all levels of transmitted drug resistance across Europe. Strong dominance of resistance to drugs which are not very popular anymore. Strong increase in transmitted NNRTI resistance and decline in resistance to protease inhibitors.
Epidemiology HIV	Data on the incidence and prevalence of HIV in Amsterdam, where the first models will be based.
Contact networks	Mathematical description for contact networks

Why are these results of benefit to DynaNets

The results will be used for parameterization of the model or used for validation. The biological parameters will be important as they determine transmission of the virus and progression through the different stages. The result that transmission of drug resistance is predominantly onward (between antiretroviral naïve patients) and not forward (from treatment experienced to newly infected) will also be included in the model. Previous modeling studies did not take this into account and may therefore have resulted in biased results. The prediction of time to virological failure will be included as viral load is the key determinant of transmission of HIV. This result is important to be included, but unlikely to make a strong difference in the predictions as onward transmission was found to be more important. The past levels of transmitted drug resistance will be used for validation. We have used the results (increase of NNRTI, strong presence of resistance to drugs that are not popular anymore) as a basis for further investigation in our case studies.

Part 2 University rankings

In part 2 we discuss the progress that was made on university collaboration networks (this is part of the DynaNets extension). In this part we created a map of links of scientific cooperation between universities in different fields. For this purpose, we determined the relation between the rank of a university (as defined by independent rankings of universities across the World) and the number of publications. We found a correlation between the number of papers with international co-authorship (i.e. more prestigious universities that are top ranked rank have more publications with international co-authorship).

In summary, our preliminary results show that even such fundamental and straightforward analysis as calculation of correlation coefficient between position of the university in the ranking and the number of papers published by its employees may reveal some non-trivial relationships. In particular, one may use it as indicator of the interest a certain scientific area gains over the years. Thus it can be possible to spot an emergence of certain trends in science and, in effect, react for example establishing a new direction of research in the university. On the other hand the analysis of common publications with an imposed threshold led to an observation of a scientific structure on the international level.

Part 3 Handling of missing data

In this part we defined different techniques for handling missing data. These techniques can be used in different settings where different variables may be missing. These techniques include the use of phylogenetic data to approximate transmission networks, treatment of missing data in human interactions and Monte Carlo filtering techniques.

Conclusion

DynaNets made strong progress in the second year regarding analysis of data. We identified several key features explaining (drug resistant) HIV infections (including late presenters, predominance of onward transmission and levels of transmitted drug resistance). These results have been used in formulation of case studies and will be used for parameterization and validation of an agent based complex networks model. The results of the case studies will be reported upon next year. Moreover, we summarized methods that can be used handling missing data.

2 Contributors

Several institutes have contributed to this deliverable.

Partner	Contribution
EMC	Writing of executive summary, introduction, conclusions. Analysis and report on data about clinical, demographic, phylogenetic and epidemiological data. Work on missing data
UCSC	Analysis and report on data about clinical, demographic, phylogenetic and epidemiological data
ISI	Contact networks. Work on missing data
UvA	Description of model and Amsterdam Cohort Study. Work on missing data
WUT	Work on University Ranking

PART 1 – Analysis of datasets available to DynaNets

1 Introduction

DynaNets will investigate the dynamics of HIV (drug resistance) using complex network analysis. In this deliverable we will report on analysis of data available to DynaNets (these data are summarized in deliverable 2.1 which was delivered at month 6). We have developed an agent based complex networks model. This model will be used to study three case studies:

1. Impact of super-spreaders on the HIV epidemic among MSM
2. Impact of earlier start of treatment on spread of HIV
3. Factors determining the constant prevalence of transmitted drug resistance despite improved virological control by treatment. Factors determining that resistance to zidovudine and stavudine remain most frequently observed in studies among antiretroviral naive patients although these drugs are hardly prescribed.

1.1 Rationale for case study 1, impact of super spreaders

A super spreader is an infectious person that spreads an infectious disease to many other people. Super spreaders can be important in the transmission of infectious diseases. For instance, there is a study that reported that a single patient was the source of a SARS outbreak including 33 individuals in a general hospital in Tianjin, China [1].

Super spreaders may cause a disproportionate proportion of new HIV infections because they have a high rate of partner change or a high viral load [2]. The latter is the key parameter driving transmission of HIV [3]. Super spreaders may be infected with a drug resistant variant and as such become a major source of transmission of drug resistance.

The exact contribution of super spreaders to transmission of (drug resistant) HIV is not known. Identifying their contribution could be important for public health as they could be specifically targeted for prevention interventions and thereby curbing the epidemic. We will use a complex model to study the impact of super spreads on the HIV (drug resistant) epidemic. In this respect, a complex model has an important advantage over “classical” SIR models that usually do not incorporate super-spreaders.

1.2 Rationale for case study 2, earlier treatment

There is a discussion ongoing about the best time to start treatment with HIV [4-7]. In recent years guidelines have tended to be more conservative and have recommended deferring treatment until $CD4^+$-counts are <350 [4, 5]. The reasons for this are the low absolute risk of

AIDS defining illnesses at higher $CD4^+$-cell counts, side effects of antiretrovirals, inconvenience of drug regimens leading to reduced adherence, and an increased risk of drug resistance. But advances in drug development have resulted in regimens that are more convenient to use with respect to pill burden, time of dosing and less side effects. Earlier initiation of treatment may therefore be preferable [4].

Several studies have found that patients starting treatment at higher CD4 counts of <500 is associated with a reduced risk on opportunistic infections and serious non-AIDS events [7-10]. Initiation of antiretroviral drugs at higher $CD4^+$-cell counts could be beneficial for the spread of HIV as patients spend less time with detectable viral loads. But earlier treatment may also lead to a higher risk for developing resistance. We will use a complex HIV transmission model to study the impact of earlier treatment on the (drug resistant) HIV epidemic.

1.3 Rationale for case study 3, spread of drug resistant HIV

Treatment of HIV can be limited by the emergence of drug resistance. Importantly, drug resistance that emerged in a patient can be transmitted to others [11-14]. Transmitted drug resistance has clinical ramifications as it is associated with virological failure in patients who receive at least one antiretroviral drug to which the virus has lost susceptibility [15].

The prevalence of drug resistance in Europe has been stable at a level of 8-10% during the last decade [11-14]. This level remained stable despite strong reductions in the proportion of patients with virological failure during the last years [16]. In addition, drug resistance among patients newly diagnosed with HIV frequently involves thymidine associated mutations (TAMs) which confer resistance to zidovudine and stavudine. Both drugs are currently not popular in treatment and not recommended in first-line treatment [17]. But zidovudine was the only drug available for treatment of HIV in the late 1980s and early 1990s. Treatment with zidovudine as only antiretroviral drug is associated with development of drug resistance [18]. It therefore seems that resistance to zidovudine emerged early in the HIV epidemic and continues to be transmitted. Nonetheless, the epidemic of transmitted drug resistant cannot be fully ascribed to TAMs [13].

We will use a complex model to study the most likely explanation why the levels of transmitted drug resistance remain stable over time and why TAMs remain the most commonly transmitted drug resistance associated mutations.

2 Computational model for simulating the spreading of HIV

For simulating the HIV epidemic among MSM we use SEECN (Simulator for Epidemic Evolution on Complex Networks) [31], an agent-based computational model where the agents are connected with a dynamic complex network. Each agent represents an individual MSM with individual infection status (e.g., healthy, acute phase, asymptomatic and treated phase) and behavior (having casual sexual encounters, steady relationships, or both).

Time is discretized into units, and in each time unit all dynamics (for both agents and the network) are applied. These dynamics include: removing or adding agents, progressing an agent's infection status (e.g., from asymptomatic to AIDS), infecting another agent over a network link, and changing the network structure. All parameters must be exhaustively specified in matrix form to reduce the chance of mistakes and to enable automation within the framework of WP5.

The model for the case studies discussed below has 10 possible statuses for agents, of which there are three types (having casual sex only, having steady relationships only, or both), and two types of network links (steady and casual). The list of possible agent statuses is as follows:

1. Healthy
2. Acutely infected (first three months after infection), but no drug resistance
3. Acutely infected, with transmitted drug resistance
4. Asymptomatic phase, no diagnosis, no treatment, no drug resistance
5. Asymptomatic phase, no diagnosis, no treatment, with drug resistance
6. Asymptomatic phase, with diagnosis, no treatment, with drug resistance
7. Asymptomatic phase, with diagnosis, no treatment, no drug resistance
8. Asymptomatic phase, with treatment, without drug resistance
9. Asymptomatic phase, with treatment, with drug resistance
10. AIDS

As input to SEECN simulations, not only single values for parameters are used. Rather, for each parameter (such as infection probability or time-to-AIDS) we specify a probability distribution based on the uncertainty of the parameter value, such as a confidence interval (CI) if reported in literature or a Poisson distribution for results based on sparse sampling (such as cohort studies). The reason for this is that it is indefensible to use only single

(maximum likelihood) values for each parameter, since the likelihood of that single value is often still close to zero and small changes in parameter values may have large impact on the predictions.

Instead, a separate program stores the probability distributions for all parameter values. It calls SEECN a large number of times, each time generating new parameter values according to the distributions and feeding these values to SEECN. The results of SEECN are stored along with the parameter values that were used for later analyses, such as expected epidemic size or the epidemic impact as function of an agent's number of edges, along with the corresponding uncertainties. These 'corresponding uncertainties' are important: they tell us to what extent any conclusion can be made at all with the current amount of data. This is lacking in most existing epidemic studies.

The reason for using SEECN is that it simulates individual agents that are connected with a dynamic network. To achieve this computationally demanding task, SEECN is optimized for performance by cache efficiency and parallelism.

3 Results

In this section we will describe the main results that were obtained by analyzing the datasets available to DynaNets. More detailed results that are needed in the complex network are summarized in Tables that are attached as an appendix.

3.1 Data related to biology of HIV infection

Several biological factors control the transmission dynamics of HIV. These factors include the transmissibility of the virus and disease progression. This information was retrieved from the literature as it was not available in the datasets available to DynaNets. We will therefore not discuss this information in detail but instead refer to the Table in chapter 5 of Part 1.

3.2 Data related to sexual networks

Number of partnerships among men-having-sex-with-men (MSM)

The Dutch Schorer foundation (See: http://www.schorer.nl/16/about-schorer/) is the Netherlands institute for homosexuality, health and well-being. Amongst others the Schorer foundation carries out prevention campaigns in the area of HIV and other sexually transmitted infections.

Schorer has collected information amongst Dutch gay men about the number of sexual partners and duration of relationships (see Table in chapter 6 for detailed information). The more than 5000 participating men reported a median of 8 sexual partners (range 3-20) during the previous six months. A third of participants reported 1-9 lifetime partners, 30% of respondents reported 11-50 partners, and almost 33% reported 51 to 500 partners. One out of 12 participating men indicated to have had >500 sexual partners during their life.

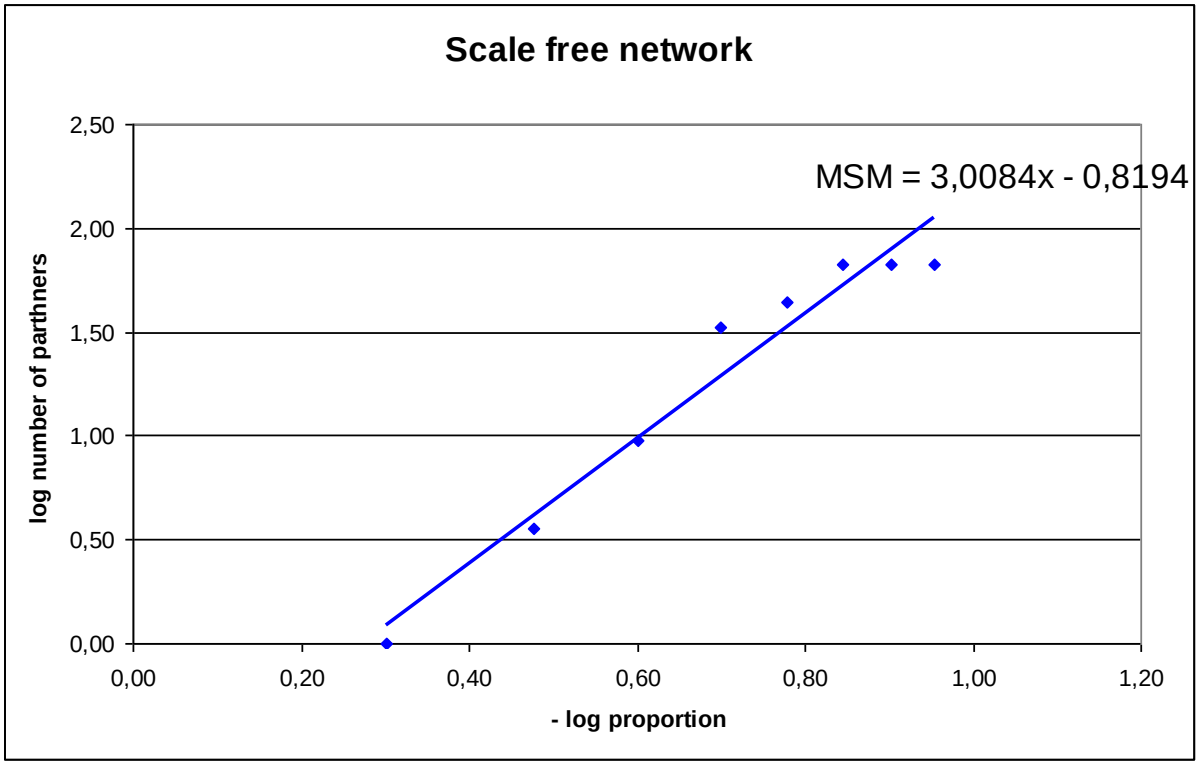
Sexual networks and HIV contact tracing

Contact tracing is a potential public health intervention in which individuals who test positive for HIV are asked about (recent) sexual partnerships. The EMC has access to data on HIV contact tracing. We will not further report on this because of the confidentiality of the data.

Phylogenetically clustered transmissions in pan-European SPREAD programme

We had access to data from the pan-European SPREAD programme. SPREAD collects information about transmission of drug resistant HIV-1 in 25 countries across Europe. Included are patients newly diagnosed with HIV-1 that are representative for the risk group and geographical distribution of HIV in the participating countries. In 2006 and 2007 a total of 1630 patients were included. We used the phylogentic tools reported in deliverable 2.2 to identify phylogenetically clustered transmissions. The graph below shows the association between the cumulative number of patients in a cluster (on a log-10 scale) versus the cumulative proportion (on a log-10 scale)

Part 1 Figure 1: Scale free network phylogenetic analysis SPREAD



A Novel Methodology for Large-Scale Phylogeny Partition and Application to the Italian HIV-1 Epidemic

(This is a short summary of a technique that was reported in deliverable 2.2. The results will be used and we therefore shortly repeat the method and main results).

Rationale: Understanding the determinants of virus transmission is a fundamental step for effective design of screening and intervention strategies to control viral epidemics. Phylogenetic analysis can be a valid approach for the identification of transmission chains, and very-large datasets can be analyzed through parallel computation.

Study design and methods: We designed and validated a new methodology for the partition of large-scale phylogenies and the inference of transmission clusters. This approach, based on a depth-first search algorithm, conjugates the evaluation of node reliability, tree topology and patristic distance analysis. The phylogenetic analysis was carried out using parallel computation, and the estimated phylogeny was linked to epidemiological and clinical data. Randomised tests for proportions and multivariable regression were used to assess factors associated with transmission chains.

Summary of principal findings: The method has been validated using epidemiologically confirmed transmission chains from an independent cohort available publicly in the HIV Los Alamos repository. Additional validation tests were performed on simulated data sets.

The method was then applied to identify transmission clusters of a phylogeny of 11,541 HIV-1 subtype B pol gene sequences from a large Italian cohort (ARCA). Molecular transmission chains were characterized by means of different clinical/demographic factors. In particular we found that transmission chains are associated with a recent infection, a high viral load and generally they are characterised by non-mixed risk groups. However, there was also a significant mixing between male homosexuals and male heterosexuals. Transmission of drug resistant HIV was largely explained by onward transmission between antiretroviral naïve patients. Transmission of resistance from patients failing treatment was rare. In a large perspective, since the method takes advantage of a flexible notion of transmission cluster, it can become a general framework to analyze other epidemics. The method has been implemented in a platform-independent application using java to be released as a free software.

3.3 Data related to clinical care of HIV infection

For our model we need to know at what CD4 count and viral load individuals are diagnosed. In case, interventions to curb epidemic could in principle be less effective when many patients are diagnosed at a relatively late stage of their HIV infection. Patients identified in a late stage may not have adapted their risk behaviour or received treatment which strongly reduces their viral load which in turn reduces their infectivity.

For the case study on the impact of earlier treatment, it is necessary to determine the time it takes for a CD4 count to go from 500 (suggested as a cut-off to start earlier treatment by some) to a cD4 count of 350 (cut-off to start treatment in current treatment guidelines). We have analyzed three different datasets available to DynaNets. The results are outlined in the following paragraphs.

Database Eramus MC, Rotterdam

The database at the Erasmus MC in Rotterdam includes data from more than 2000 HIV-infected patients that are treated for their HIV infection.

The figure below shows the distribution in CD4 count for 2010 patients at entry at the Erasmus MC. Excluded were almost 200 patients that were already on treatment. The figure includes four different CD4 categories:

1) <200 which is used as one of the criteria of defining a person as having AIDS,
2) between 200 and 350. These patients can be defined as late presenters as they should have started treatment
3) between 350 and 500. These patients would be eligible for earlier treatment
4) >500, these patients are not eligible for treatment

Figure 2 shows that a substantial proportion of patients only came under medical attention at an advanced stage of infection.

We also determined the time to start treatment in each CD4 stratum. Patients that had a CD4 between 350 and 500 started treatment a median of 1.75 years after entry into care. During this time MSM could have a substantial number of sexual partners (section 5.1.2) and earlier treatment may therefore have a beneficial impact on the HIV epidemic.

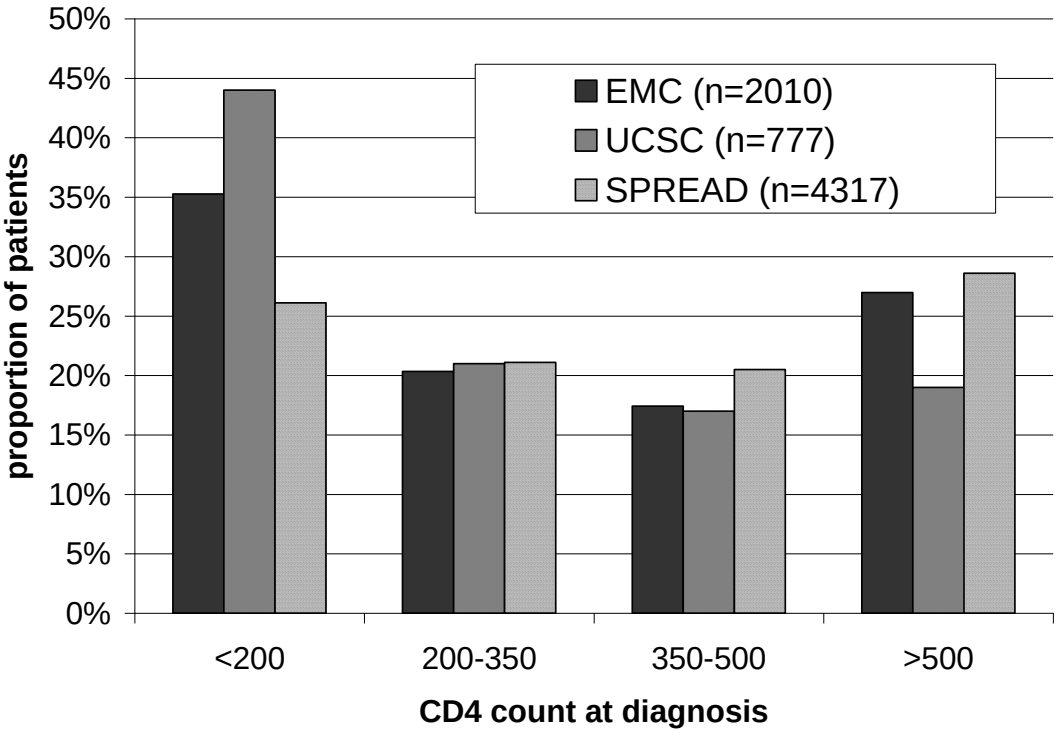
Pan-European SPREAD database

SPREAD is a cross-sectional dataset on transmission of drug resistance. As a consequence, it is not possible to determine the time to a particular event in the participating patients.

The 1630 patients included in SPREAD in the years 2006 and 2007 had a median CD4 count of 370 (inter-quartile range 210 to 548). Almost one out of three patients were infected for <1 year at time of diagnosis. Importantly, a considerable proportion of patients were at an advanced stage of infection as indicated by a proportion of 11.5% of patients that were diagnosed in CDC stage C.

The CD4 count at time of diagnosis in 4317 patients included in the SPREAD study is shown in Figure 2.

Part 1 Figure 2: Distribution of CD4 count at diagnosis



Database UCSC, Rome

We found 777 patients (out of a total of 4388 enrolled at UCSC) with a known first HIV positive test (average year was 2003), a CD4+ T cell count measurement within 30 days from the first positive test, and at least one subsequent CD4+ T cell measurement thereafter. Of these patients, 34% were female, 72% of Italian nationality, 47% heterosexual, 8% IDU, 20% homosexual, 22% with a known date of HIV sero-negativity, the average year of birth was 1965, 50% were on CDC stage A, 16% CDC stage B, 30% CDC stage C, 3% of patients died afterwards. The average (st.dev) CD4+ T cell count corresponding to the first HIV positive test date was 295 (255) cells/mm3. Figure 2 shows the distribution of CD4 count at time of diagnosis.

The proportion of patients who started an antiretroviral therapy (ART) afterwards was 86%. The time to start treatment for patients diagnosed with a CD4 between 350 and 500 was 686 days. This number is comparable with the 645 days found for this group of patients in the database at ErasmusMC (Table).

A Prognostic Model for Estimating the Time to Virologic Failure in HIV-1 Infected Patients Undergoing a New Combination Antiretroviral Therapy Regimen

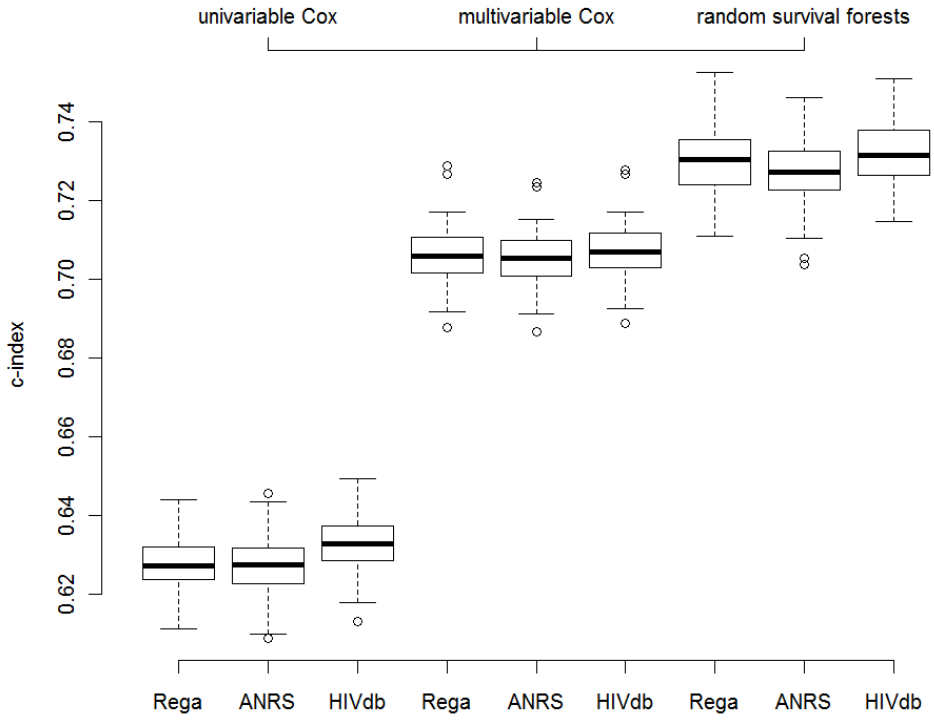
Rationale: HIV-1 genotypic susceptibility scores (GSSs) were proven to be significant prognostic factors of fixed time-point virologic outcomes after combination antiretroviral

therapy (cART) switch/initiation. However, their relative-hazard for the time to virologic failure has not been thoroughly investigated, and an expert system that is able to predict how long a new cART regimen will last has never been designed.

Study design and methods: We analyzed patients of the Italian ARCA cohort starting a new cART from 1999 onwards either after virologic failure or as treatment-naïve. Endpoint was the time to virologic failure, from the 90th day after treatment start, defined as the first HIV-1 RNA >400 copies/ml, censoring at last available HIV-1 RNA before treatment discontinuation. We assessed the relative hazard/importance of GSSs according to distinct interpretation systems (Rega, ANRS and HIVdb) and other covariates by means of Cox regression and random survival forests (RSF). Prediction models were validated via the bootstrap and c-index measure.

Summary of principal findings: The dataset included 2337 regimens from 2182 patients, of which 733 were previously treatment-naïve. We observed 1067 virologic failures over 2820 persons-years. Multivariable analysis revealed that low GSSs of cART were independently associated with the hazard of a virologic failure, along with several other covariates, as illustrated in Table 1. Evaluation of predictive performance yielded a modest ability of the Cox regression to predict the virologic endpoint (c-index≈0.70), whilst RSF showed a better performance (c-index≈0.73, $p<0.0001$ vs. the Cox), as illustrated in Figure 3. Variable importance according to RSF was concordant with the Cox hazards. In conclusion, GSSs of cART and several other covariates were investigated using linear and non-linear survival analysis. RSF are a promising approach for the development of a reliable system that predicts time to virologic failure better than Cox regression. Such models might represent a substantial improvement over the current state-of-the-art methods for monitoring and optimization of cART. In DynaNets, the results will be used to predict the time to virological failure. This information is important as patients who fail virologically can spread their HIV infection to others.

Part 1 Figure 3: Extra-sample performance evaluation of Cox regression and Random Survival Forests models by means of c-index distributions (100 bootstrap runs). Boxplots indicate average and interquartile range, whilst whiskers indicate 95% confidence intervals.



3.4 Data related to drug resistance

Pan-European SPREAD dataset

The table below (Table 1) shows the distribution of transmitted drug resistance found in the SPREAD dataset.

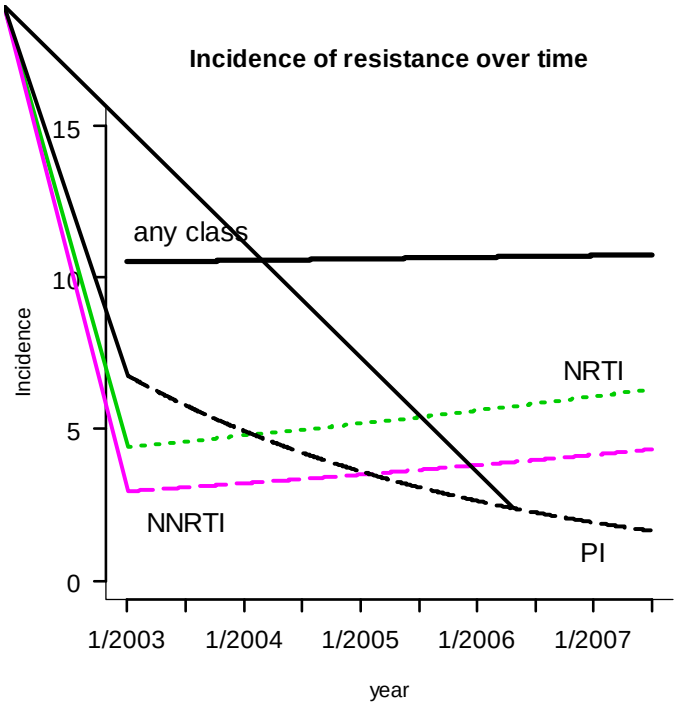
Part 1 Table 1: Prevalence of transmission of drug resistant HIV-1 in the SPREAD study

	Number (%)
At least one resistance associated mutation	158 (9.7)
NRTI drugs class	93 (5.7)
NNRTI drugs class	63 (3.9)
PI drugs class	27 (1.7)
Resistance to one class	137 (86.7)
Multi Drug Resistance (≥2 classes)	17 (10.8)
Multi Drug Resistance (≥3 classes)	4 (2.5)

Resistance was for the largest part limited to a single class of antiretroviral drugs.

We also determined the trend over time by comparing the results from the current collection-period (2006-2007) to previous periods (2002-2005). The result of the time trend analysis is provided below in Figure 4. Please note that the results are expressed as incidence meaning that we limited the analysis to patients that were infected for <1 year.

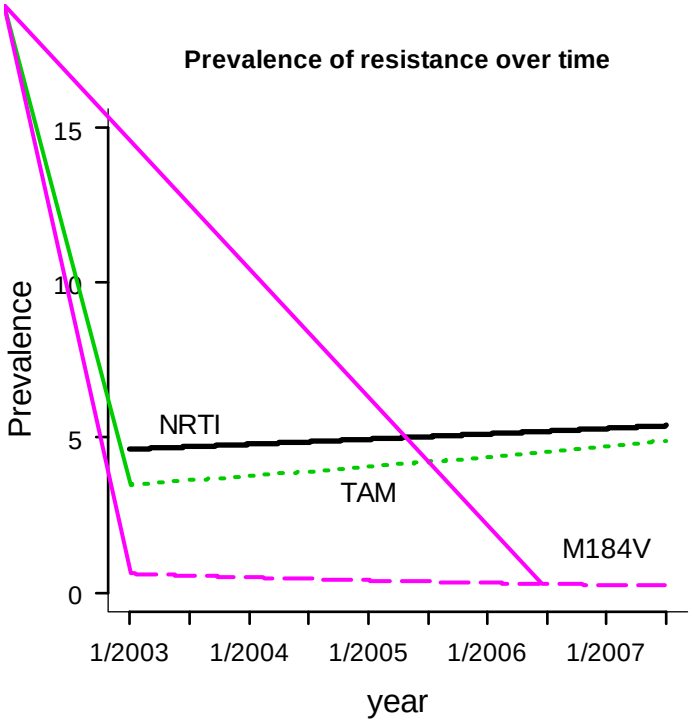
Part 1 Figure 4: Incidence of transmitted drug resistance over time across Europe



Of note is that the prevalence of transmitted resistance remains fairly stable over time. Resistance to particular classes of antiertrovirals does, however, change. Resistance to NRTIs and NNRTIs is increasing. A strong decrease in transmitted resistance to protease inhibitors (PIs) is observed. This change could possibly be ascribed to the increaesed popularity of boosted protease inhbitors which have a high genetic barrier for drug resistance making the emergence of drug resistance less likely [19].

We also determined time trends regarding NRTI-resistance. We paid particular interest to TAMs and to the lamivudine resistance associated mutation M184V. This mutation is the most frequently observed mutation in patients failing treatment across Europe [20, 21]. Importantly, transmitted resistance to NRTIs is for the largest part ascribed to TAMs. The M184V is only observed in a limited number of patients.

Part 1 Figure 5: Prevalence transmitted NRTI drug resistance over time across Europe



It is important to know if the pre-dominance of TAMs in transmission of drug resistant can be explained by current treatment guidelines. Table 2 is a summary of the international treatment guidelines that are published every two years by the IAS (International AIDS Society). Please note that in recent years the thymidine analogues zidovudine and stavudine are not recommended as first line treatment (they were recommended up until 2002). This suggests that transmission from patients failing treatment is not the main driver for transmission of resistance to newly diagnosed patients. This observation is supported by phylogenetic data from ARCA which showed that drug resistance seems predominantly the consequence of onward transmission between antiretroviral naïve patients and not from a patient failing treatment to a newly infected individual (see section 3.2).

Part 1 Table 2: Summary of recommended first line regimens over time

Year	1st line therapy	alternative
2010	efavirenz or a ritonavir-boosted PI (atazanavir; darunavir, raltegravir) plus 2 NRTI (tenofovir/emtricitabine or abacavir/lamivudine)	Ritonavir-boosted lopinavir or fosamplrenavir, or maraviroc
2008	efavirenz or a ritonavir-boosted PI (lopinavir, atazanavir, fosamprenavir, darunavir, saquinavir) plus 2 NRTI (tenofovir/emtricitabine or abacavir/lamivudine)	
2006	efavirenz or a ritonavir-boosted PI (lopinavir, atazanavir, fosamprenavir, or saquinavir) plus 2 NRTI tenofovir/emtricitabine, zidovudine/lamivudine, or abacavir/lamivudine.	
2002	(indinavir, nelfinavir, ritonavir)/saquinavir; ritonavir/indinavir; ritonavir/lopinavir; or efavirenz plus 2 NRTI (didanosine/lamivudine; stavudine/didanosine; stavudine/lamibudine; zidovudine/didanosine; zidovudine/lamivudine)	
1998	2 NRTIs (zidovudine/didanosine; stavudine/didanosine; zidovudine/zalcitabine; zidovudine/lamivudine; stavudine/lamivudine) and 1 PI (indinavir, nelfinavir, ritonavir, saquinavir)	ritonavir and saquinavir (with one or two NRTIs) or nevirapine

3.5 Amsterdam Cohort Study

The data from the Amsterdam Cohort Study (ACS) were used as one of the basis of the model we used for the case study. The ACS of HIV infection and AIDS among MSM living in Amsterdam was initiated in 1984 and has involved 2299 men until 2006 [22]. Of the 2299 MSM, 571 were HIV-positive at study entry and 192 seroconverted during follow-up until the end of 2006. Consequently, characteristics of MSM's behavior and surveyed yearly HIV-incidence over calendar years for the population are reported periodically [23].

HIV-positives are seen every three months. Clinical, epidemiological and social scientific data are collected with standardised questionnaires (six monthly) and by physical examination. Blood is taken for virological and immunological tests and for storage. HIV-negatives are seen by a nurse every six months and similar data are collected but no immunological tests are done nor are cells stored. The surival rate and cause of death is actively assessed once a year.

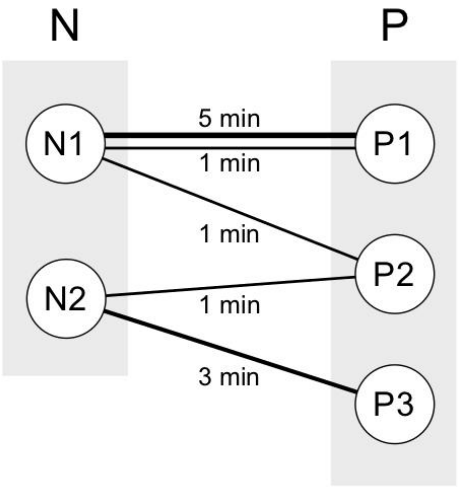
The main contribution of ACS to our studies is the incidence and prevalence data of HIV of the various stages.

3.6 Contact networks

Mathematical modeling of disease spreading within a population typically relies on so-called contact matrices [1-8] characterizing the interaction between individuals belonging to different classes.

The traditional approach to data collection through interviews and recall of previous encounters is not based on objective measurements and is generally performed on a random day, thus lacking the longitudinal dimension [9].To overcome such limitations we rely on RFID technology as described in [10–15]. We stress that the mathematical description reported here can also be applied to networks of sexual contacts partitioned, for instance, into different age groups. For each pair of individuals i and j we define several possible weights w_{ij} , each corresponding to a different quantity measured on the collected data: the occurrence of the contact w_{ij}^{p} , with $w_{ij}^{p} = \mathbf{1}$ if at least one contact between i and j has been established, and 0 otherwise; the frequency of the contact w_{ij}^{n} , indicating how many times the contact between i and j is observed during the study; the time spent on each such encounter; the cumulative duration of the contact w_{ij}^{t} , indicating the sum of the durations of all contacts established between i and j observed during the study. In addition to the above quantities, that are weights defined for pairs of individuals i and j, it is possible to define corresponding quantities s_i for each individual i, aggregating on all individuals j who had a contact with i. In relation to the previously defined weights, one obtains the following quantities: the number of distinct contacts s_i^{p} ,indicating the number of distinct individuals with whom i has established at least one contact (i.e., a contact between i and j that occurs $w_{ij}^{n} > \mathbf{1}$ times is counted only once); the number of contacts s_i^{n}, indicating the overall number of contacts established by individual i, counting repeated contacts with the same individual j as distinct events; the cumulative time in contact s_i^{t} , corresponding to the total sum of the duration of all contacts involving individual i. Figure 6 provides an example on how the above quantities are computed for a schematic sequence of contact data, where the individuals i and j belong to two different classes, nurses (N) and patients (P).The collected data allow thus to inspect the interaction behavior by focusing on face-to-face proximity between pairs of categories.

Part 1 Figure 6: Schematic representation of detected contacts among 2 nurses (N1, N2) and 3 patients (P1, P2, P3) and corresponding measured quantities. Each individual is represented by a node and a link corresponds to a contact established between two individuals. The width of the link is a measure of the duration of the contact, also indicated explicitly in terms of minutes. Multiple links can occur between two individuals, as highlighted in the pair N1-P1, indicating a contact of frequency larger than 1. The pair established one contact ($w_{ij}^{p} = \mathbf{1}$) with frequency equal to two (w_{ij}^{n}) for a total duration of six minutes ($w_{ij}^{t} = \mathbf{6}$ min). By taking into account all interactions, individual N1 has established three contacts ($s_i^{n} = \mathbf{3}$), two of which were distinct contacts ($s_i^{p} = \mathbf{2}$), for a total duration of contacts equal to seven minutes ($s_i^{t} = \mathbf{7}$ min).



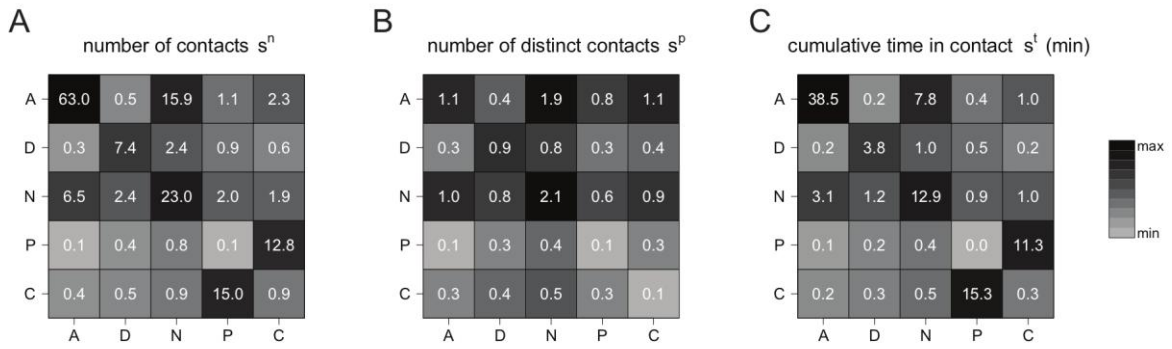
$\mathbf{w^{p}}_{N1,P1} = 1$ $\mathbf{s^{p}}_{N1} = 2$

$\mathbf{w^{n}}_{N1,P1} = 2$ $\mathbf{s^{n}}_{N1} = 3$

$\mathbf{w^{t}}_{N1,P1} = 6$ min $\mathbf{s^{t}}_{N1} = 7$ min

Figure 7 reports three contact matrices defined on the classes of participants (ward assistants (A), nurses (N), physicians (D), patients (P) and caregivers (C)), taking into account the different numbers of individuals in each class [3], and measured using the three quantities defined above: the number of contacts s_i^{n} (panel A), the number of distinct contacts s_i^{p} (panel B), and the cumulative time in contact s_i^{t} (panel C).

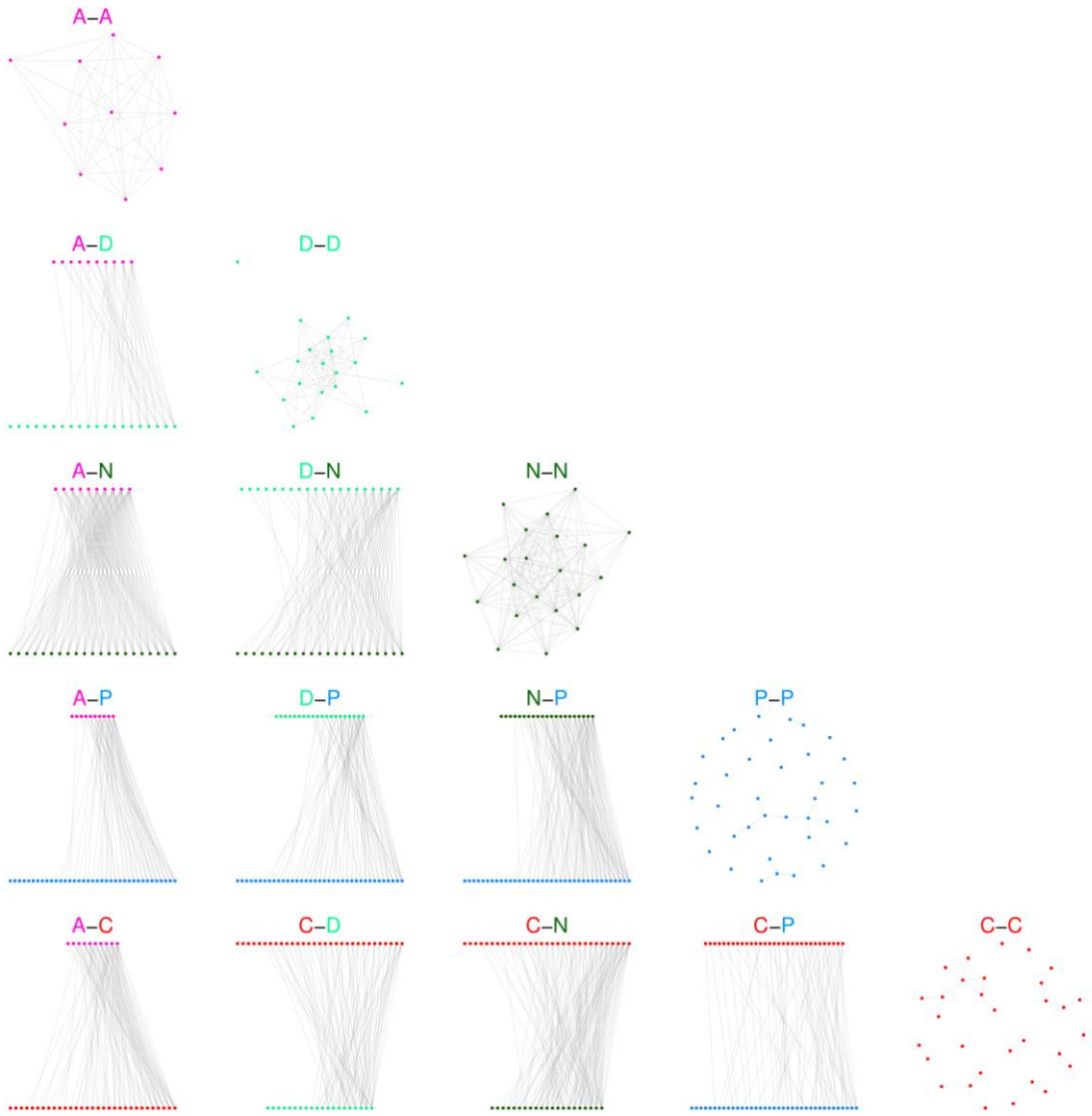
Part 1 Figure 7: Contact matrices defined on the classes of individuals. Matrices are displayed for the number of contacts (panel A), the number of distinct contacts (panel B), and the cumulative time in contact (panel C). The matrix entry for classes X (row) and Y (column) is the median value of the node strengths for individuals of class X, computed on the contacts they had with individuals of class Y; the asymmetry of the matrices depends on the different numbers of individuals populating each class [3]. Individuals of class X that did not have contacts with individuals of class Y count as nodes with zero strength, i.e., they affect the median value for the corresponding matrix entry. To increase the readability of the figure, matrix entries are grayscale-coded according to the median values, with the lightest and darkest shade of gray respectively corresponding to the minimum and maximum value for each matrix. Contact durations are expressed in minutes and normalized to a 24-hour interval.



Panel A shows that the majority of contacts occur within the ward assistant class, followed by nurse-nurse interactions. A number of contacts larger than 10 is also observed for patient-caregiver interactions, considering both the number of contacts that a patient had with any caregiver, and the number of contacts that a caregiver had with any patient. These contacts, however, are characterized by a very high frequency, as signaled by the very small number of distinct contacts reported in panel B for the same two classes, consistent with a strong one-to-one patient-caregiver interaction. A smaller number of distinct contacts is also observed among ward assistants, compared to the median value of 63 contacts, whereas the nurse-nurse interaction remains strong also in terms of number of distinct contacts. The number of contacts (both distinct and non-distinct) among patients is noticeably very low and close to zero. The contact matrix computed in terms of the cumulative time in contact provides yet another characterization of the interaction behavior among classes. Long interactions are observed between patients and caregivers, among nurses, and among ward assistants. The time spent in close proximity by a pair of patients or by a pair of visitors is extremely small, and interactions between a health care worker and a non-health care worker are very limited. Figure 8 shows a set of interaction networks for each pair of classes

and within each class, corresponding to the entire monitoring period. In these networks, a node represents an individual, and an edge is drawn between two individuals whenever a face-to-face proximity event involving them was recorded. The networks restricted to physicians, nurses or ward assistants are rather dense, indicating a large diversity of contacts. On the other hand there are only very few contacts between caregivers or between patients, but, as expected, the patient-caregiver contacts are very specific. Each patient has contacts with essentially one caregiver, and vice-versa. Contacts among caregivers and among patients are barely observed.

Part 1 Figure 8: Cumulative contact networks of individuals, for all pairs of classes and within each class. Nodes represent unique individuals, and edges between nodes represent a cumulative face-to-face time over the whole monitoring period. In the off-diagonal layouts, nodes are positioned from left to right in increasing order of number of edges.



4 Conclusion

DynaNets made strong progress in the second year regarding analysis of data. We identified several key features explaining (drug resistant) HIV infections (including late presenters, predominance of onward transmission and levels of transmitted drug resistance). These results have been used in formulation of case studies and will be used for parameterization and validation of an agent based complex networks model. The results of the case studies will be reported upon next year.

5 Tables with detailed results

Part 1 Table 3: Detailed results data analysis

Parameter		*Additional description*	*Value, range*	*Dataset/Reference*
Data related to biology of HIV infection				
Transmissibility of HIV		*MSM[1], per contact*		
		Insertive UAI[2]	0.62% (95% CI 0.07-1.68)	[22]
		Receptive UAI with withdrawal	0.65% (0.15-1.53)	[22]
		Receptive UAI with ejaculation	1.43% (0.48-2.48)	[22]
		Heterosexual, transmission hazard per year, per partnership		
		Acute stage, first 0.24 year	2.76 (95% CI 1.31-5.09)	[23, 24]
		Chronic stage, 8.38 years	0.106 (0.0761-0.133)	[23, 24]
		AIDS stage, 0.75 years	0.76 (0.413-1.28)	[23, 24]
		Final AIDS stage to death	0	[23, 24]
		Reduction transmissibility by treatment	92%-100%	[25-27]
Disease progression		Time from seroconversion to AIDS and death	Detailed description in reference	[28]
Data related to sexual partners				
Number of partners		Number of life time sexual partners	N=5476	Schorer monitor, NL
		1-10	29%	
		11-50	30%	
		51-500	32%	
		>500	8%	

Parameter		*Additional description*	*Value, range*	*Dataset/Reference*
		Last six months	Median 8 (range 3-20)	
Data related to clinical care of patients				
Viral load at entry		Log 10	EMC, UCSC, SPREAD Median: 4.8 / 4.5 / 4.8 IQR: 4.2-5.4 / 3.9-5.1 / 4.1-5.0	ErasmusMC, UCSC, SPREAD
Duration of infection		<1 year 1-2 years Unknown duration	511 (31.3%) 68 (4.2%) 1051 (64.5%)	SPREAD
CD4 count at diagnosis		Figure 2 <200 200-350 350-500 >500	EMC/UCSC/SPREAD 35% / 44% / 26% 20% / 21% / 21% 17% / 17% / 21% 27% / 19% / 29%	ErasmusMC, UCSC, SPREAD
Time to treatment		CD4 at entry <200 200-350 350-500 >500	ErasmusMC, UCSC Median (IQR)[3] days to treatment 38 (22-81) 118 (43-398) 643 (197-1226) / 686 (387-869) 1170 (623-2041)	ErasmusMC, UCSC

[1] MSM = Men-having-Sex-with-Men

[2] UAI = Unprotected Anal Intercourse

[3] IQR = Inter-quartile Range

Part 1 Table 4Table 4 detailed results multivariable Cox proportional hazard model showing relative hazards (RH) for time-to-virologic-failure, fitted on the whole study population (n=2,337).

Factor		*RH*	*95% CI*	*p-value*
calendar year	before 2004 vs. 2007 and after	2.06	(1.67-2.54)	<0.0001
	2004 vs. after 2007 and after	1.62	(1.29-2.03)	<0.0001
	2005-2006 vs. 2007 and after	1.28	(1.06-1.55)	0.0109
cART	2NRTI+1PI vs. 2NRTI+1NNRTI	1.03	(0.83-1.27)	0.8028
	2NRTI+1PI/r vs. 2NRTI+1NNRTI	0.63	(0.54-0.75)	<0.0001
	3NRTI vs. 2NRTI+1NNRTI	1.23	(0.92-1.64)	0.1599
age (per 10 years older)		0.89	(0.82-0.96)	0.0036
gender (male vs. female)		1.06	(0.91-1.23)	0.4668
mode of HIV-1 transmission	male homosexual vs. heterosexual	1.08	(0.88-1.33)	0.4680
	IDU vs. heterosexual	1.08	(0.87-1.32)	0.4898
	other/unknown vs. heterosexual	1.08	(0.9-1.29)	0.4248
nationality	non-Italian vs. Italian	1.23	(0.9-1.67)	0.1992
	unknown vs. Italian	0.94	(0.78-1.14)	0.5475
HCV/HBV coinfection	unknown vs. no	1.18	(0.97-1.45)	0.1049
	yes vs. no	1.04	(0.82-1.32)	0.7277
HIV-1 RNA per log10 copies/ml higher		1.27	(1.17-1.39)	<0.0001
CD4+ count cells/mm3	<=100 vs. >500	1.57	(1.23-2)	0.0003
	>100 and <=199 vs. >500	1.16	(0.93-1.45)	0.1968
	>200 and <=349 vs. >500	1.22	(1-1.48)	0.0447
	>350 and <=499 vs. >500	0.98	(0.79-1.21)	0.8202
interval time from the first HIV-1 positive test to ART initiation	<=12 vs. >60 months	0.87	(0.67-1.13)	0.2944
	>12 and <=60 vs. >60 months	1.01	(0.81-1.27)	0.9114
	unknown vs. >60 months	0.92	(0.77-1.11)	0.3938
duration of prior ART exposures	<=6 vs. >24 months	0.84	(0.7-1.01)	0.0626
	>6 and <=12 vs. >24 months	0.92	(0.73-1.17)	0.5029
	>12 and <=24 vs. >24 months	0.83	(0.66-1.03)	0.0890
previous AIDS-defining events (yes vs. no)		0.86	(0.7-1.05)	0.1379
#previous ART switches		1.03	(1-1.05)	0.0522
previous ART class exposure	NRTI vs. ART-naïve	1.48	(1.01-2.17)	0.0441
	NRTI and NNRTI vs. ART-naïve	1.38	(0.99-1.93)	0.0546
	NRTI and NNRTI and PI vs. ART-naïve	1.43	(1.03-1.99)	0.0315
	NRTI and NNRTI and PI/r vs. ART-naïve	2.96	(2.16-4.06)	<0.0001
	NRTI and PI vs. ART-naïve	2.18	(1.64-2.89)	<0.0001
	NRTI and PI/r vs. ART-naïve	2.72	(1.98-3.75)	<0.0001
	other classes vs. ART-naïve	2.31	(1.31-4.05)	0.0036

previous exposure to suboptimal ART (yes vs. no)		0.85	(0.71-1.03)	0.0946
viral subtype	02_AG vs. B	0.98	(0.61-1.57)	0.9402
	C vs. B	1.41	(0.86-2.32)	0.1748
	F1 vs. B	0.57	(0.3-1.06)	0.0750
	other vs. B	1.26	(0.76-2.09)	0.3774
	undetermined vs. B	1.12	(0.88-1.42)	0.3527
GSS*	ANRS per 1 point increase	0.72	(0.66-0.78)	<0.0001
	HIVdb per 1 point increase	0.68	(0.63-0.74)	<0.0001
	Rega per 1 point increase	0.71	(0.66-0.77)	<0.0001

RH: relative hazard; CI: confidence interval; cART: combination antiretroviral therapy; NRTI: nucleoside/nucleotide reverse transcriptase inhibitors; NNRTI: non-nucloeside reverse transcriptase inhibitors; PI: protease inhibitors; PI/r: ritonavir-boosted PI; IDU: injecting drug users; HCV: hepatitis C virus; HBV: hepatitis B virus; ART: antiretroviral therapy; GSS: genotypic susceptibility score; *fitted separately one from each other.

PART 2 – University ranking

1 Objectives

The objective of this part of WP2 (DynaNets extension) is to provide data about University collaboration network as well as: Data Warehousing, Analysis and Network Design.

2 Introduction

The role of the WUT in this WP is to work within Project Area A3 *Network centric data mining* by data acquisition and corresponding analysis for University collaboration networks. Our analysis is bound to measure correlations between a node degree in such a network and University ranks received from University ranking lists. The objective is to check whether there are research fields bringing the links that decide on the University position.

The schematic plan of the activities within the Deliverable is following:

- Identification of datasets on the collaboration of scientists (such as INSPEC, Science Citation Index, arXiv:
- Acquisition of datasets regarding the rank of the universities, colleges and other research institutes being the employers of the above mentioned scientists.
- Data cleansing using existing methods or methods currently developed in the DynaNets project.
- Extraction of correlations between the rank of the institution and the intensity of the collaboration of its employees for specific research field (physics, mathematics, chemistry etc.).

It is believed that the studies might be of help in search for an improvement of management strategies at EU Academic / Educational Institutions.

3 Aims

We process the analysis by creating a map of links of scientific cooperation between universities in different fields. The key issue to be answered can be formed as: do the best universities collaborate with the best ones only? We anticipate that after ordering the papers according to publication years, we will observe changes of interest in different fields in a sense that one discipline could disappear and others could emerge and be correlated with universities ranks.

4 Data sources

In order to estimate the correlations between university rankings and scientific productivity we had to identify two different sources of data:

a) first devoted to the university ranking with at least 10 years activity
b) second connected to actual bibliographic information, in particular complying with the following rules:
    - allowing to view categories of publications,
    - allowing to view address of the publication,
    - allowing to view year of publication.

(a) The lists of top hundred universities were downloaded from two services: *www.arwu.org* (Academic Ranking of World Universities – later referred to as ARWU) and *www.topuniversities.com* (QS World University Ranking – later referred to as QS). The rationale behind choosing two rankings that follow different rules was to check the robustness of the performed analysis.

(b) After preliminary analysis, we have chosen the service Web of Science as a data source for obtaining the information on citations. For one institution the average number of publications ranges between few to dozens of thousands of publications. As a result each university has two tables containing the following fields:

- published (date of publication)
- ID (reference to the second table)
- subject category (category of publications)
- language

The key information used in this report is the subject category of the published paper. Below we give examples of several of them (the total number is 180):

Part 2 Table 1: Examples of subject categories in the Web of Science database.

Biology	Medicine	Cardiac and Cardiovascular Systems	Biochemistry and Molecular Biology
Neurosciences	Physics	Cell Biology	Public
Immunology	Oncology	Chemistry	Hematology
Surgery	Multidisciplinary	Psychiatry	Peripheral Vascular

	Sciences		Disease
Radiology	Psychology	Clinical Neurology	Microbiology

5 Data verification

5.1 Abbreviations

The seemingly straightforward procedure of querying for a specific university name encounters some problems that could have a strong impact on the further results. Web of Science has a set of abbreviations commonly used for searching such as *Univ* for "University" or *Coll* for "College". Moreover it is essential to notice that one has to form a very specific query in order to get rid of severe mistakes. Below in Table 2 we show an exemplary list of the search universities together with the exact search phrase that had to be used.

Part 2 Table 2: Universities names and queries giving results obtained for given universities

Rank	University	Search	Country
1	Harvard University	Harvard Univ	United States
2	University of Cambridge	Univ Cambridge	United Kingdom
3	Yale University	Yale Univ	United States
4	UCL University College London	UCL	United Kingdom
5	Imperial College London	London Imperial Coll	United Kingdom
5	University of Oxford	Univ Oxford	United Kingdom
7	University of Chicago	Univ Chicago	United States
8	Princeton University	Princeton	United States
9	Massachusetts Institute of Technology MIT	MIT	United States
10	California Institute of Technology Caltech	Caltech	United States
17	Australian National University	Australian Natl Univ	Australia
67	London School of Economics and Political Science LSE	LONDON SCH ECON and POLIT SCI	United Kingdom
67	Lund University	Lund Univ	Sweden
69	KAIST - Korea Advanced Institute of Science amp Technology	Korea Adv Inst Sci and Technol	Korea, South
70	Utrecht University	Univ Utrecht	Netherlands

70	University of York	Univ York	United Kingdom
72	University of Geneva	Univ Geneva	Switzerland
73	Nanyang Technological University NTU	Nanyang Technol Univ	Singapore
73	Washington University in St Louis	Washington Univ+St Louis	United States
78	University of North Carolina, Chapel Hill	Univ N Carolina+Chapel Hill	United States
98	Ludwig-Maximilians-Universität München	Univ Munich|Tech Univ Munich	Germany

5.2 Ambiguity of queries

The 'Search' field is a search key that we use to associate with the authors of the publications and it can consist of one of the operators:

- "+" which stand for *and* operator in Boolean logic
- "|" which stand for *not* operator in Boolean logic

These operators are used to clearly assess the origin of the publication. Table 2 shows that using just the names of universities from the list (first column) would lead in the case of No. 98 to obtaining publications of both Technical University in Munich and University of Munich, instead of just the latter. To omit this problem one has to insert a query *Univ Munich*|*Tech Univ Munich* that ensures achieving proper results. On the other hand for the case shown as no. 78, it was not sufficient to enter *Washington Univ*, as there are many universities with such an abbreviation; it was necessary to add *St. Louis* in the query text.

5.3 Size limitations

We discovered that only the publications from the years 2004 – 2010 are reliable due to a limitation imposed by the maximum query to the database.

6 Methodology

(a) The easiest and quickest way to spot in a qualitative way the relation between the rank of universities *R* and number of publications *P* in the given subject category is to plot *P(R)* on a 2D graph. Those are presented in the next section.

(b) To estimate in the quantitatively the level of correlations between the rank of the university *R* and the number of papers *P* published in the given subject area we used Pearson's correlation coefficient defined as

$$r_{RP} = \frac{\sum_{i=1}^{n}\left(R_i - \overline{R}\right)\left(P_i - \overline{P}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_i - \overline{R}\right)^2 \sum_{i=1}^{n}\left(P_i - \overline{P}\right)^2}} \quad . \qquad (1)$$

The coefficient has a following interpretation: if two variables *R* and *P* tend to be highly correlated *r* is close to 1, while for totally anti-correlated variables it reaches -1. In case the variables are not correlated at all the coefficient *r* is equal to 0.

It is important to notice here that the variable *P* can mean:

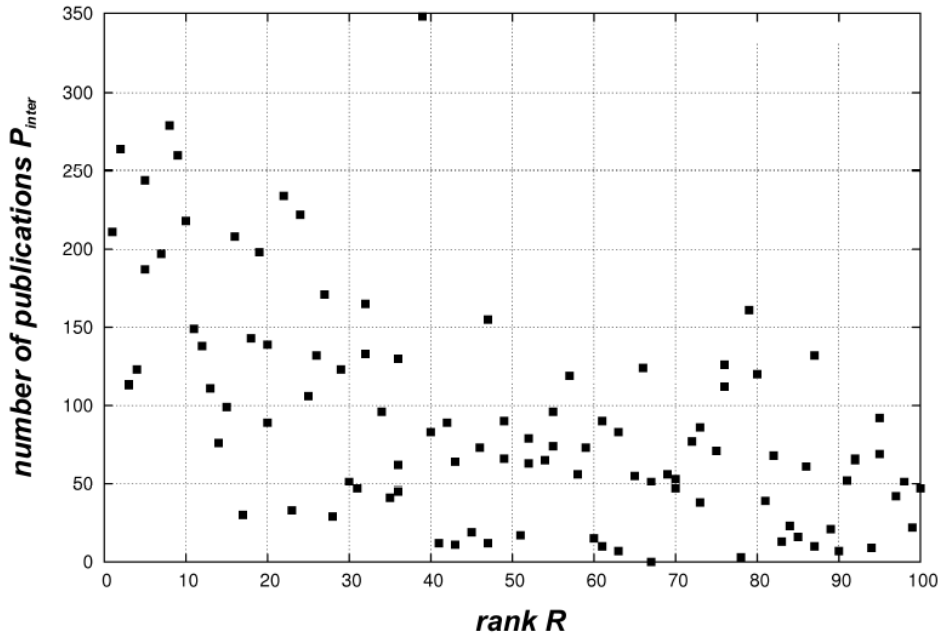
- the number of all papers P_{all},
- the number of individual papers (i.e., papers that have only one author) P_{ind},
- the number of papers of which all the authors are employees of the given university P_{univ},
- the number of papers with the national co-authorship P_{nat},
- the number of papers with the international co-authorship P_{inter}.

(c) One can also analyze the direct connections between universities *i* and *j* on the basis of the collaboration matrix ***C*** where the element C_{ij} gives the number of common publications of institutions *i* and *j*.

7 Results

Below we show the dependence between the number of papers with the international co-authorship P_{inter} and the university rank *R* in the case of subject category Arts. The plot is made for the ARWU ranking for the years 2009-2010. One immediately notices that large number of publications is connected to high position in the ranking list (i.e., a low number *R*). It is confirmed by the value of the correlation coefficient *r=-0.58* suggesting a significant anti-correlation between P_{inter} and *R* in this category.

Part 2 Figure 1: **Number of publications with international co-authorship P_{inter} versus university rank *R* (taken from the ARWU ranking) in the years 2009-2010.**



In fact, as is shown in Table 3, subject category Arts has the highest correlation value in the case of ARWU ranking. On the other hand, while considering the QS ranking, it is the Multidisciplinary Sciences that holds the highest place with *r* even higher than in the case of Arts.

Part 2 Table 3: Correlations coefficients r between the rank of university *R* and the number of papers with international co-authorship P_{inter} for different subject categories in years 2009-2010. The table on the left-hand side takes the rank from ARWU while the one on the right-hand side – from QS.

ARWU	
Subject category	**r**
Art	-0,58
Economics	-0,50
Physics	-0,50
Social Sciences	-0,50
all	-0,50
Multidisciplinary Sciences	-0,50
Developmental Biology	-0,50
Statistics and Probability	-0,49
Biochemistry and Molecular Biology	-0,49
Genetics and Heredity	-0,48

QS	
Subject category	**r**
Multidisciplinary Sciences	-0,63
Art	-0,60
Chemistry	-0,58
Biophysics	-0,58
Physics	-0,56
all	-0,55
Biochemical Research Methods	-0,55
Economics	-0,55
Humanities	-0,53
Biochemistry and Molecular Biology	-0,53

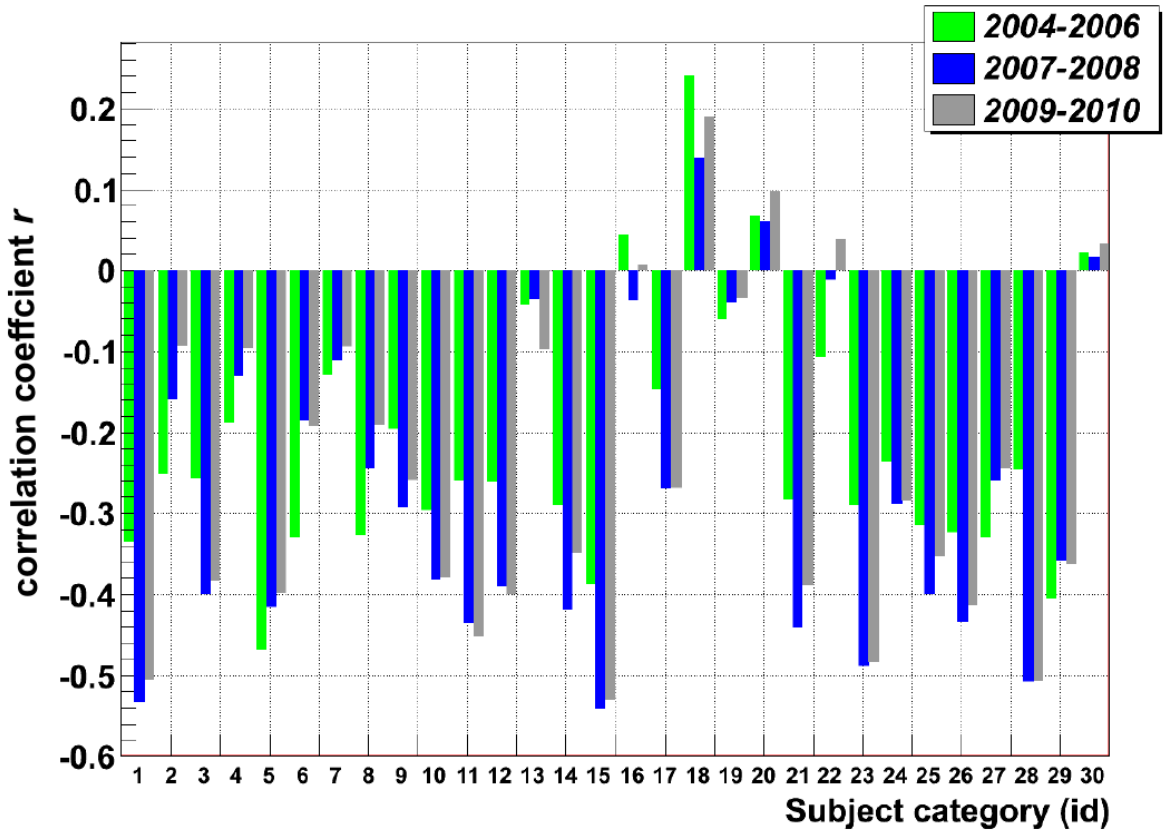
The temporal range of the downloaded data gives the opportunity to follow the evolution of the value of correlation coefficient r in different years. In this way one can compare if in a given subject category there is an improvement (or not) of the correlation over the period of time. For example Fig. 2 clearly shows that (ARWU ranking, all publications) in the consecutive years correlation coefficient *r* gradually drops down for the subject categories Physics (Id=4), Biophysics (Id=2) and Computer Science (Id=7). That might suggest a decreasing impact of those areas on the position a university holds in the ranking. On the other hand, such area as Sociology (Id=28) doubled its correlation over the years 2004-2008.

Using Pajek program (www.pajek.org) it is possible to visualize connections between universities that are the outcome of the matrix C mentioned in section 6(c). However, one has to impose a threshold on the number of common publications in order to be able to observe any structure. Figure 3 depicts a network where each node is a university and a link between two nodes is drawn if two universities have at least 500 common papers. There is a clear sign large number links related to the high-ranked universities (upper left corner of the plot), while the institutions placed in the end of the ranking do not tend to have almost any of them.

Imposing even higher threshold (1000 common papers) and color-coding the countries the universities belong to leads to even more interesting results: one identifies certain universities that serve as kind of “bridges” between countries (Fig. 4). The prominent role is

played here by the ETH Zurich which appears to be a “scientific hub” between Europe and America.

Part 2 Figure 2: **Correlation coefficient r between university rank R and the total number of all publications Pall for different subject categories (see Table 4 for specific names) in three periods: 2004-2006 (green), 2007-2008 (blue) and 2009-2010 (grey).**

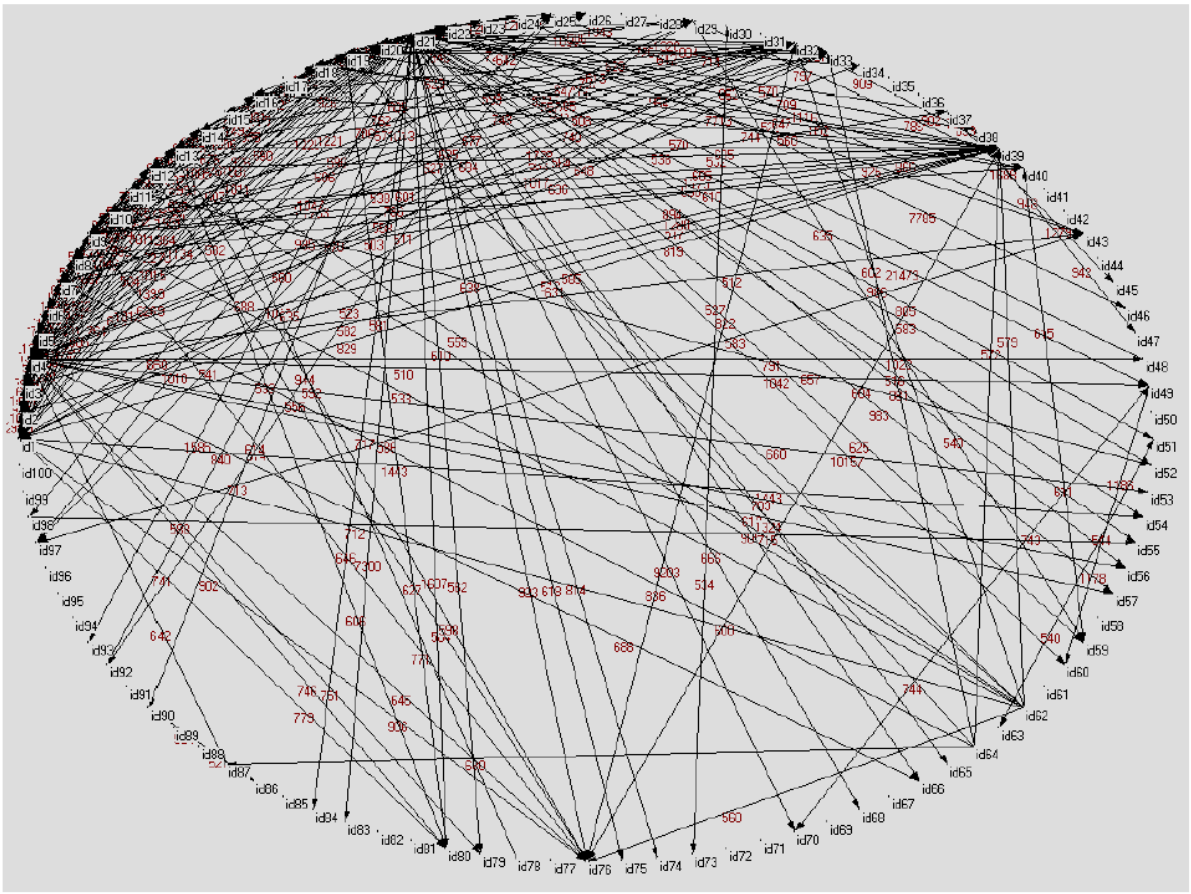


Part 2 Table 4: Names of subject categories Figure 2

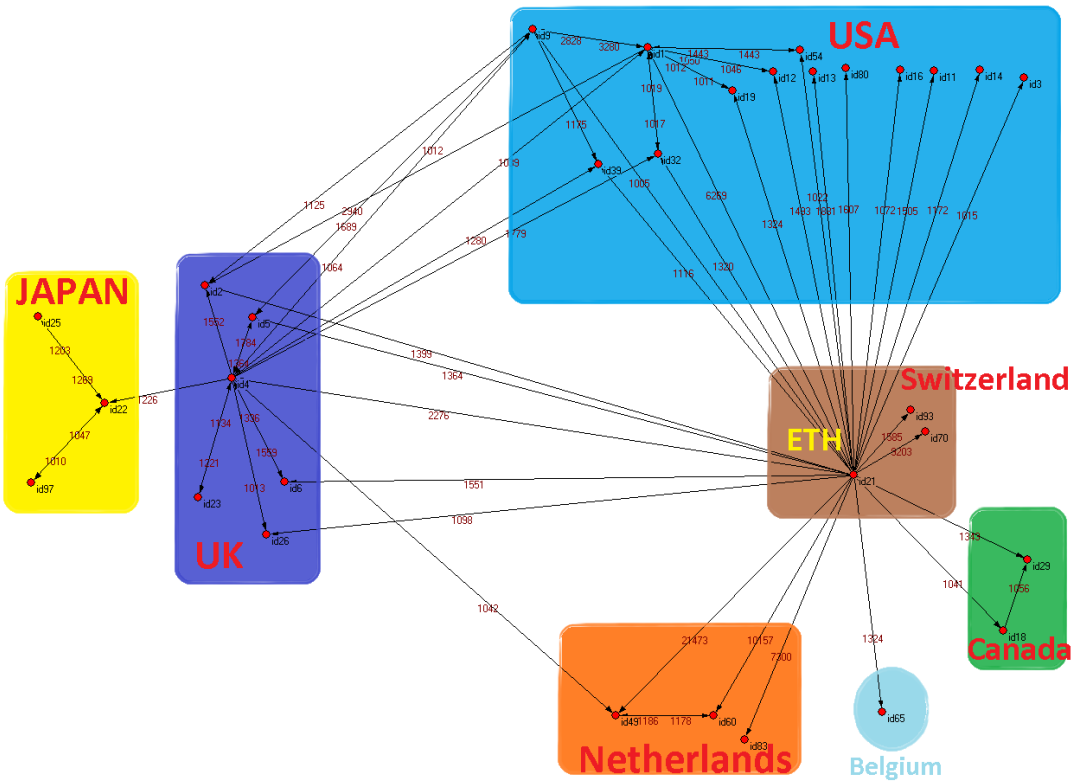
Id	Subject category	Id	Subject category
1	Acoustics	16	Nutrition and Dietetics
2	Biophysics	17	Agricultural Engineering
3	Engineering	18	Energy and Fuels
4	Physics	19	Agriculture
5	Imaging Science and Photographic Technology	20	Behavioral Sciences
6	Communication	21	Veterinary Sciences
7	Computer Science	22	Genetics and Heredity
8	Language and Linguistics	23	Reproductive Biology

9	Radiology	24	Chemistry
10	Neuroimaging	25	Ecology
11	Obstetrics and Gynecology	26	Environmental Sciences
12	Agricultural Economics and Policy	27	History and Philosophy Of Science
13	Biotechnology and Applied Microbiology	28	Sociology
14	Economics	29	Agronomy
15	Food Science and Technology	30	Forestry

Part 2 Figure 3: Network of Universities with at least 500 common papers



Part 2 Figure 4: Network of universities with at least 1000 common papers



8 Conclusion

Our preliminary results show that even such fundamental and straightforward analysis as calculation of correlation coefficient between position of the university in the ranking and the number of papers published by its employees may reveal some non-trivial relationships. In particular, one may use it as indicator of the interest a certain scientific area gains over the years. Thus it can be possible to spot an emergence of certain trends in science and, in effect, react for example establishing a new direction of research in the university. On the other hand the analysis of common publications with an imposed threshold led to an observation of a scientific structure on the international level.

PART 3 – Handling missing data

1 Introduction

The description of work of DynaNets for the first year stated that deliverable 2.2 would study methods for handling missing data. The work on missing data was, however, not reported in deliverable 2.2. The review report of year 1 stated that handling missing data is important. We will therefore discuss these methods in this deliverable (as reported in our comments to the review report after year 1).

1.1 Phylogenetic data to approximate transmission network

Available data: individual behavior data, clinical data, cohort studies
Missing data: network structure of sexual contacts over which HIV spreads
In order to effectively model and simulate the HIV epidemic using agent-based modeling, extensive data is required at various spatiotemporal scales. We will use three levels: virological, individual, and population. At the virological level, we have 'sufficient' data about phenomena such as the time to development of drug resistance, acute phase duration, time to AIDS, and reduction of infectiousness due to treatment. By 'sufficient' we mean that we can estimate the parameters within acceptable bounds; within these bounds we will employ the sampling techniques Metropolis and Monte-Carlo in order to take the remainder uncertainty into account. At the individual (behavioral) level, we also have sufficient data on phenomena such as the frequency of condom use (both in casual encounters and in steady relationships), the duration of a steady relationship, and the number of sexual contacts in a steady relationship.

What is missing, however, is sufficient data on the population level – i.e., the network structure formed by the sexual contacts integrated over some time. This is required as input to the simulations to decide 'who can infect who' at what time. The current best guess is that the network structure is approximately scale-free with exponent between 1.5 and 2.0 for MSM when integrating contacts over a year which comes from literature and is based on sparse data. We expect that translating this fuzzy description of the network structure to our simulations, where time steps are three months (instead of a year) and the population is different in size (Rotterdam, Amsterdam) and slightly different in culture and geography, the remainder uncertainty that the simulations would have to sample from and average over would become too large. If this happens then not only becomes the computation prohibitive, also the averaged quantities (such as predicted fraction of drug resistant patients) are not representative anymore. The latter happens because two significantly different network structures could give significantly different predictions, and yet both are possible given the available data.

Our plan to overcome this missing data is to use phylogenetic data as a likelihood function for the network structure. The phylogenetic data form an independent observation of the HIV epidemic among MSM and can be used to decide which network structures are more likely, i.e., reproduce the phylogenetic data more faithfully. In concrete terms, the phylogenetic data can be translated to a cluster-size distribution, where a 'cluster' is a group of individuals which have infected each other (in-)directly. This genome-independent statistic can also be measured in the simulated HIV epidemics among MSM, each time assuming a particular network structure, which can be compared with the cluster-size distribution of the phylogenetic data to decide how likely that network structure is. Of course, for each network structure we have to do multiple simulations because most other parameters also have uncertainties to be sampled from (Metropolis), as well as because of stochasticity (Monte-Carlo).

Concluding, the network structure of sexual contacts is quite uncertain because of missing data on 'who had sex with who', but this can be partly remedied by using additional information from phylogenetic data which forms an independent observation of how the epidemic spreads.

1.2 Human interactions

*Data and infrastructure*The experimental framework aims at measuring the contact patterns of a group of interacting individuals in a spatially bounded setting, such as a set of offices or a conference. The participants are asked to carry small RFID tags [22], henceforth called beacons. These beacons continuously broadcast small data packets which are received by a number of stations and relayed through a local network to a server. The stations are installed at fixed locations in the environment. The beacons and stations we used were created by and obtained from the OpenBeacon project [23].

RFID tags acting as beacons can be used to deploy indoors locative systems [24] that track the location of the tags. Problems related to multiple path, phase fluctuations, etc. tend however to limit the precision of the spatial localization of the tags. Because of this, locative technologies are typically not viable, at low cost, to infer face-to-face contact between individuals wearing RFID tags.

Moving from contact inference to direct contact detection enabled us to bypass these limitations. To this end, we leveraged the OpenBeacon active RFID platform [23] and operate the RFID tags a bi-directional fashion. That is, tags no longer act as simple beacons that

passively emit signals to be received and processed by a centralized post-processing setup. They rather exchange messages in a peer-to-peer fashion to sense their neighbourhood and assess directly contacts with nearby tags.

A high spatial resolution of less then 1 - 2 meters is attained by using very low radio power levels for the contact sensing. Furthermore, assuming that the subjects wear the tags on their chest, the body effectively acts as a shield for the sensing signals. This way, contacts are detected only when participants actually face one another. If a sensed contact persists for a few seconds, then given the short range and the face-to-face requirement, it is reasonable to assume that the experiment is able to detect an ongoing social contact (as e.g. a conversation).

After the beacons detect a contact, they broadcast a report message at a higher power level. These reports are received by the stations and relayed to the monitoring infrastructure. The reports are stored with a time stamp, the id of the relaying station and the id of the tags which participate in the contact event (up to 4 simultaneous contacts are recorded, using the current hardware). [22] Finkenzeller, K.: RFID handbook, Wiley Hoboken, NJ (2003). [23] OpenBeacon project, http://www.openbeacon.org/.

Treatment of missing data The aggregated networks of human interactions are empirical networks i.e. the information about nodes and edges in the networks is obtained via a direct measurement. Unlike the case of the HIV spreading simulations, one does not know the key parameters driving the dynamics of the system. As a consequence of the absence of an analytical model, the maximization of the likelihood function is not applicable to these empirical networks. Most of the noise in the collected data comes from human mishandling of the RFID badges or on occasional hardware failures.

Among the former there is the case of wearable RFID devices, returned by the participants to the data collection campaign, which were not switched off. Such an event would lead to the appearance of an unphysical persistent cluster of interacting individuals. The latter may include devices whose battery is discharging, thus leading to an excessive number of reboots or to long contact durations ending up broken into many shortly-sustained contacts. Malfunctioning antennas may lead to a similar scenario. Due to the absence of a model for dynamics, we resort to a non-parametric sensitivity analysis of the noise. Below we sketch the various steps we follow.

(1) The first step consists in calculating a set of quantities of interest on the original data set.
(2) We then assess the amount of noise in the collected data. We deploy computer scripts to automatically process the experimental data and flag potentially corrupted data (PCD).
(3) The quantities of interest are then recalculated without the PCD. Such a procedure is of course viable only when we have a vast amount of collected data or the amount of noise in the collected data is almost negligible, as it may discard valuable information.
(4) A more refined approach tries to fix directly the PCD. For instance, many intermittent contacts between two individuals may be replaced by a single long contact and clusters of interacting individuals persisting for many hours are disregarded in the analysis. Once again we calculate the quantities of interest on this "cured" data set.
(5) Finally, we may add some artificial noise is the system by rewiring and/or removing at random a small fraction of links in the aggregated network. Once again, the evaluation of the quantities of interest on this artificially-corrupted data set allows researchers to get some insight on the robustness of the statistics for the quantities of interest.
(6) The estimated quantities of interest in points (1-5) are compared and conclusions about the quality of the statistical indicators are drawn.
Despite having been developed for a specific case study (networks of human interactions), variations of the procedures described in points (1-6) can be applied to any empirical network.

1.3 Monte Carlo filtering techniques

For some studies, there is information available about the size of the epidemic at a particular moment in time. Key information about sexual partnerships may, however, be missing. For example, the antenatal HIV prevalence is well documented in most parts of Africa. Data about partnerships could be (partially) missing. In this case, Monte Carlo filtering techniques can be used. In these techniques, a large number of simulations is performed. In the end, only those simulations that fall within well-defined limits will be accepted and used for further analysis of the model.

References

1. Wang Sh X, Li YM, Sun BC, et al. The SARS outbreak in a general hospital in Tianjin, China -- the case of super-spreader. Epidemiol Infect 2006;134:786-91
2. Chen L, Jha P, Stirling B, et al. Sexual risk factors for HIV infection in early and advanced HIV epidemics in sub-Saharan Africa: systematic overview of 68 epidemiological studies. PLoS One 2007;2:e1001

3. Quinn TC, Wawer MJ, Sewankambo N, et al. Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. N.Engl.J.Med. 2000;342:921-929
4. Hughes MD, Ribaudo HR. The search for data on when to start treatment for HIV infection. J Infect Dis 2008;197:1084-6
5. Hammer SM, Saag MS, Schechter M, et al. Treatment for adult HIV infection: 2006 recommendations of the International AIDS Society-USA panel. JAMA 2006;296:827-843
6. Strategies for Management of Antiretroviral Therapy Study G, Lundgren JD, Babiker A, et al. Inferior clinical outcome of the CD4+ cell count-guided antiretroviral treatment interruption strategy in the SMART study: role of CD4+ Cell counts and HIV RNA levels during follow-up. J Infect Dis 2008;197:1145-55
7. Strategies for Management of Antiretroviral Therapy Study G, Emery S, Neuhaus JA, et al. Major clinical outcomes in antiretroviral therapy (ART)-naive participants and in those not receiving ART at baseline in the SMART study. J Infect Dis 2008;197:1133-44
8. Prosperi MC, Cozzi-Lepri A, Castagna A, et al. Incidence of malignancies in HIV-infected patients and prognostic role of current CD4 cell count: evidence from a large Italian cohort study. Clin Infect Dis 2010;50:1316-21
9. Ho JE, Deeks SG, Hecht FM, et al. Initiation of antiretroviral therapy at higher nadir CD4+ T-cell counts is associated with reduced arterial stiffness in HIV-infected individuals. Aids 2010;24:1897-905
10. Kitahata MM, Gange SJ, Abraham AG, et al. Effect of early versus deferred antiretroviral therapy for HIV on survival. N Engl J Med 2009;360:1815-26
11. Spread-programme. Transmission of drug-resistant HIV-1 in Europe remains limited to single classes. Aids 2008;22:625-35
12. van de Vijver DAMC, Wensing AMJ, Boucher CAB, et al. The epidemiology of transmission of drug resistant HIV-1. HIV Sequence Compendium 2006/2007: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. LA-UR 07-4826, 2007:17-36
13. Vercauteren J, Wensing AM, van de Vijver DA, et al. Transmission of drug-resistant HIV-1 is stabilizing in Europe. J Infect Dis 2009;200:1503-8
14. Wensing AM, van de Vijver DA, Angarano G, et al. Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management. J.Infect.Dis. 2005;192:958-966
15. Wittkop L, Gunthard HF, de Wolf F, et al. Effect of transmitted drug resistance on virological and immunological response to initial combination antiretroviral therapy for HIV (EuroCoord-CHAIN joint project): a European multicohort study. Lancet Infect Dis 2011

16. Gill VS, Lima VD, Zhang W, et al. Improved virological outcomes in British Columbia concomitant with decreasing incidence of HIV type 1 drug resistance detection. Clin Infect Dis 2010;50:98-105
17. Thompson MA, Aberg JA, Cahn P, et al. Antiretroviral treatment of adult HIV infection: 2010 recommendations of the International AIDS Society-USA panel. Jama 2010;304:321-33
18. Boucher CA, O'Sullivan E, Mulder JW, et al. Ordered appearance of zidovudine resistance mutations during treatment of 18 human immunodeficiency virus-positive subjects. J.Infect.Dis. 1992;165:105-110
19. van de Vijver DA, Wensing AM, Angarano G, et al. The calculated genetic barrier for antiretroviral drug resistance substitutions is largely similar for different HIV-1 subtypes. J.Acquir.Immune.Defic.Syndr. 2006;41:352-360
20. Frentz D, Boucher CA, Assel M, et al. Comparison of HIV-1 genotypic resistance test interpretation systems in predicting virological outcomes over time. PLoS One 2010;5:e11505
21. van de Vijver DA, Wensing AM, Asjo B, et al. HIV-1 drug-resistance patterns among patients on failing treatment in a large number of European countries. Acta Dermatovenerol Alp Panonica Adriat 2010;19:3-9
22. Jin F, Jansson J, Law M, et al. Per-contact probability of HIV transmission in homosexual men in Sydney in the era of HAART. Aids 2010;24:907-13
23. Hollingsworth TD, Anderson RM and Fraser C. HIV-1 transmission, by stage of infection. J Infect Dis 2008;198:687-93
24. Wawer MJ, Gray RH, Sewankambo NK, et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. J Infect Dis 2005;191:1403-9
25. Boily MC, Buve A and Baggaley RF. HIV transmission in serodiscordant heterosexual couples. Bmj 2010;340:c2449
26. Del Romero J, Castilla J, Hernando V, Rodriguez C and Garcia S. Combined antiretroviral treatment and heterosexual transmission of HIV-1: cross sectional and prospective cohort study. Bmj 2010;340:c2205
27. Donnell D, Baeten JM, Kiarie J, et al. Heterosexual HIV-1 transmission after initiation of antiretroviral therapy: a prospective cohort analysis. Lancet 2010;375:2092-8
28. Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis. Collaborative Group on AIDS Incubation and HIV Survival including the CASCADE EU Concerted Action. Concerted Action on SeroConversion to AIDS and Death in Europe. Lancet 2000;355:1131-7