IST-2004-027173

Eu*Resist*

Integration of viral genomics with clinical data
to predict response to anti-HIV treatment

Instrument: STREP
Thematic Priority 2: “Information Society Technologies”

D2.1 - Standard Datum

Due date of deliverable: 15/08/06
Actual submission date: 23/08/06

Start date of project:1 Jan ‘06 Duration: 30 months

Organisation name of lead contractor for this deliverable: UniSiena
Authors: Maurizio Zazzi
Main contributions: Anders Sonnenborg, Rolf Kaiser
Revision [Final]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	**x**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of Contents

List of Figures

List of Tables

Executive summary

This document describes the Standard Datum as defined in the Project. The Standard Datum is the set of attributes to be used to extract a large number of records from the integrated database deemed as suitable for training the machine learning methods to be used in the Project for prediction of response to antiretroviral treatment.

Two different definitions have been developed, both based on updated knowledge on HIV dynamics during antiretroviral treatment and structure of medical records typically available from the clinic.

The first definition, hereafter referred to as the Classical Standard Datum, has been derived from the basic indications available from recent literature or discussed at recent meetings on HIV drug resistance. In its simplest structure, the standard datum is built around a treatment start or change and made of the kind of treatment itself, some baseline parameters, to be modelled as predictors, and some follow-up parameters, to be modelled as indicators of the effectiveness of therapy. However, the array of parameters commonly included in the definition has been considerably expanded in EuResist with the aim of testing the contribution of additional variables to the response to treatment and thus improving its prediction through modelling.

The second definition, the so-called Alternative Standard Datum, was originally developed by one of the Project partners (MPI) facing the issue of a limited amount of available clinical data. This standard datum is based on a minimum requirement of data, being built just on genotype-treatment pairs labelled as success or failure using a simple but clinically relevant definition. Some models preliminarily trained through this simplified kind of standard datum have shown promise and thus the definition has been retained, with minor modifications, in the context of the much larger EuResist integrated database.

Both the Classical and the Alternative Standard Datum will be used to generate training tables for the different modelling techniques considered in EuResist. In addition to working on these "summary" tables, the whole integrated database is available on-line to all the model developers of the EuResist team for further exploring possible ways to train their learning methods with the full complement of variables and the complete data set.

1. Introduction

The "standard datum" is the set of variables that are extracted from the database and pooled together to generate the typical record describing the event to be studied by mathematical modelling. The event the EuResist Project specifically focuses on is the in vivo response to a new anti-HIV treatment regimen as measured by defined indicators and in the presence of defined baseline parameters. The generation of a table made of standard datum records is thus an extraction and simplification of information from a much larger and more complex database in order to generate the appropriate data set for modelling the event of interest.

Based on a lively and prolonged discussion between individual partners and among all partners during plenary meetings, two types of Standard Datum have been defined: the so called "Classical Standard Datum" and the "Alternative Standard Datum".

Since the choice of which parameters are included in the Standard Datum has very relevant implications on the performance of the prediction engine, the two modes have been considered to cover (i) the possibility to include a larger set of variables with respect to commonly used methods (Classical Standard Datum) and (ii) the possibility to explore a simplified description of the event of interest resulting in a much larger amount of records (Alternative Standard Datum).

The Classical definition is built around the so-called "treatment change episode" (TCE), defined as the start of a new (first or subsequent to a previous one) treatment regimen (http://www.hivforum.org/uploads/Resistance/DataAnalysisPlanRev1.pdf; http://www.hivrdi.org/log/ax.pl?/pdf/RDI-DataGuidelines-1.pdf). To be included as a usable record, each TCE must be accompanied by a minimum of baseline as well as follow-up data. Typically, the baseline parameters have been just the virus genotype and the viral load obtained within a defined time frame (usually 0 to 1-3 months before the start of treatment) whereas the response to treatment has been measured as the change in viral load at short term (usually around 3 months after treatment). While the variation in viral load is a well accepted surrogate marker of the effectiveness of treatment, the virus-host interplay is fairly complex and keeping the definition of the TCE that simple may miss important variables both at baseline and follow-up (Brun-Vezinet et al., Antivir Ther. 2004;9:465-78; Miller et al., AIDS. 2006;20:929-31). For example, in light of the virus capability to archive permanently most, if not all, of the variants evolved during the course of infection (Delobel et al., AIDS 2005;19:1739-50), we have considered as baseline variables a measure of the available HIV genotypes obtained in the past. Similarly, the previously administered treatments, weighted for their duration and remoteness, have been included as surrogate for likely developed resistance and cross-resistance. On the other hand, CD4 counts have been included both as baseline parameters and as follow-up indicators of the effectiveness of the treatment, based on the frequently shown impact of baseline CD4 levels on response to treatment and on their critical role in the clinical course of the infection (Hammer et al., JAMA 2006;296:827-43).

The Classical Standard Datum proposed by other initiatives has been using only in vivo parameters since it has been inherently structured to describe an in vivo event. However, the EuResist team has also been greatly involved in generation and computation of in vitro data which may serve as surrogate estimates for in vivo drug activity. These parameters can be estimated from HIV genotype via validated functions and include: (i) the predicted level of resistance in vitro, as calculated by the support vector machine-derived geno2pheno algorithm (www.genafor.org) (Beerenwinkel et al., Proc Natl Acad Sci U S A. 2002;99:8271-6) as well as by a multiple linear regression approach, (ii) the genetic barrier to resistance, defined as the probability that the virus will not escape from drug pressure by developing (further) mutations (Beerenwinkel et al., J Infect Dis.

2005;191:1953-60). In addition, a relative effectiveness look-up table has been created to correct the predicted activity of individual drugs estimated by these approaches.

The structure of the Alternative Standard Datum has been elaborated on the basis of the most restrictive definition of success of treatment, i. e. obtaining an undetectable viral load. Any available genotype followed by a new treatment started within a defined time frame and followed by success, provides the information that the mutational pattern detected did not hinder the treatment used. On the other hand, any genotype obtained during treatment is an indication that the mutational pattern detected prevented that treatment from achieving success, since genotyping is by definition obtained from a measurable amount of circulating virus (usually >1000 copies/ml of HIV RNA). Building an Alternative Standard Datum record is thus independent of known values of baseline and follow-up viral loads measured at defined intervals: only genotype and treatment are needed and the time elapsed from treatment start to achievement of success (undetectable viral load) can be used to correct the effectiveness of the treatment, given that genotype. The obvious advantage of the Alternative Standard Datum is that a much larger number of records is generated from the same data set with respect to the Classical Standard Datum, likely at the expense of the inclusion of some possibly misleading cases. However, the Alternative Standard Datum does not take into account many of the possibly relevant variables. On the other hand, it mimics quite faithfully the clinical scenario where genotyping is requested on treatment failure, success is marked by achieving an undetectable viral load and the same treatment can first result in success (achievement of undetectable viral load) and then in failure (new genotypic assay performed).

While both the Classical and Alternative Standard Datum provide a kind of summary table whose records will be used to train the different models considered in EuResist, the EuResist integrated database is being periodically updated from the three source databases located in Germany, Italy and Sweden and made available on-line to all the EuResist teams involved in data modelling. This will let them the possibility to expand the training data set to include the full complement of available variables.

2. The Classical Standard Datum

The Classical Standard Datum is based on a "treatment change episode" and is made of mandatory as well as optional baseline and follow-up data. The inclusion of optional data has been agreed upon based on the possibility that these have an impact on response and that the different modelling strategies can deal with and may benefit from incomplete data.

Mandatory	**Baseline**	**New treatment**. Any change in composition of therapy marks a new treatment. Change means that at least one different compound has been added or withdrawn. E. g. if drugs A and B are started on 01/01/2006 and drug C is added on 04/01/2006, we mark regimen A+B start on 01/01/2006, stop on 03/01/2006; regimen A+B+C start on 04/01/2006. Changes in dosage are not considered with the exception of ritonavir (the 800-1200 mg/day full-dose vs. the 100-600 mg/day boosting dose), i. e. each ritonavir-boosted protease inhibitor is considered different from its unboosted counterpart.
		New treatment start date.
		New treatment type. Logical field with values "first" or "switch" to discriminate the very first treatment on drug-naïve patient from subsequent treatments on pre-treated patients. A patient must have completed at least two weeks of the same continuous treatment to be labeled as pre-treated.
		Baseline viral load. The value closest to the start of the new treatment among those available within 90 days before the start of the new treatment, provided there are no additional treatments lasting >14 days between the ongoing treatment at the time of viral load measurement and the new treatment (see Figure 1). Use of foscarnet (a drug used to treat CMV infection but also causing reduction of HIV RNA load) recorded in the same time frame also makes the baseline viral load invalid since the viral load at the start of the new antiretroviral treatment may be considerably lower than that measured before using foscarnet.
		Date of baseline viral load.
		Baseline genotype. The genotype (minimal PR: aa 10-95; minimal RT: aa 41-219) closest to the start of the new treatment among those available within 90 days before the start of the new treatment, provided there are no additional treatments lasting >14 days between the ongoing treatment at the time of viral load measurement and the new treatment.
		Date of baseline genotype.
	Follow-up	**Short-term follow-up viral load**. The value closest to 56 days after the start of the new and continued treatment among those available within 28-84 days after the start of the new treatment.

Optional	**Baseline**	**Reason for switch.** This should store values for "virological failure", "toxicity", "other reasons", "unknown" plus null. Particularly, this field is intended to flag cases where the treatment switch was due to toxicity since in these cases the effectiveness of the treatment may be blunt by its premature termination.
		Pre-therapy viral load. The latest HIV RNA value before initiation of the very first treatment. Cases where the viral load is reliably recorded as measured during acute infection should be discarded since the viral load may represent the transient peak rather than the steady state.
		Baseline CD4 count. The value closest to the start of the new treatment among those available within 90 days before the start of the new treatment.
		Date of baseline CD4 count.
		Baseline CD4 percentage. The value closest to the start of the new treatment among those available within 90 days before the start of the new treatment.
		Date of baseline CD4 percentage.
		Combined viral load and CD4 count. Calculated using the paired values (dates within 14 days) closest to the start of the new treatment among those available within 90 days before the start of the new treatment. The proposed function to be used is: [HIV_RNA / CD4]
		Gender.
		Age.
		Race.
		Subtype. As derived from the BLAST/phylogenetic algorithm chosen.
		Risk group. Groups considered: homosexual male, heterosexual, intravenous drug abuse, blood or blood products, mother-to-child, others.
		Past treatments (categorical mode). Each drug as category (yes or no) based on cumulative use of >3 months. Full-dose ritonavir (800-1200 mg/day) and boosting (or "baby-dose") ritonavir (100-600 mg/day) are considered two different drugs, each ritonavir-boosted protease inhibitor is considered different from its unboosted counterpart.
		Past treatments (weighted mode). Each drug as a continuous value accounting for cumulative use and time elapsed since last use. Any drug which has been cumulatively used for >3 months will have a numerical value calculated as [months_of_cumulative_use – (months_since_last_use)/10] This function implies that the effect of a previously used drug on HIV genome fades when the time elapsed since its last use exceeds tenfold the time of cumulative use in the past.
		Suboptimal treatment. Categorical (yes or no). Yes if there has been exposure to a regimen made of one NRTI (i. e. AZT DDI DDC D4T 3TC ABC TDF FTC) for >3 months or made of 2 NRTIs including 3TC/FTC for >3 months or made of 2 NRTIs not including 3TC/FTC for >6 months.
		Cumulative genotype. If a vector with 0-1 values for each relevant mutation is built on the mandatory baseline genotype, past genotypes are also accounted for by scoring any past mutation as it were present in the baseline genotype irrespective of when detected. Sometimes referred to as "historical genotype".
		Past genotypes (weighted mode). If a vector with 0-1 values for each relevant mutation is built on the mandatory baseline genotype, past genotypes are accounted for by scoring a past mutation not present in the baseline genotype as [1 / (years_since_detected)] If a mutation is present in the baseline genotype (value = 1), its previous absence should not be computed at all.

		Adherence. The record should be excluded when there is an explicit indication of poor adherence (e. g. in the field storing the reason for changing treatment in ARCA). Most of these cases have been recorded based on the physician judgment, thus there is no definition of "poor" adherence at the moment.
		Phenotype (geno2pheno mode). Each drug is scored with its predicted fold-change in susceptibility with respect to the wild type virus.
		Phenotype (MLR mode). Each drug is scored with its predicted fold-change in susceptibility with respect to the wild type virus.
		Genetic barrier. Each individual drug is scored with its predicted probability not to lose efficacy due to resistance developing in the baseline genotype.
		Relative potency. A lookup table to be used to correct the predicted activity of a drug regimen in the output (see Table 1).
		Occurrence of AIDS-defining events. A categorical variable holding true if one or more of the coded AIDS-defining events have ever occurred in the patient history.
		Number of past treatment lines. Any drug change in treatment makes a treatment line. Changes in dosage do not mark a new treatment (e. g. twice daily to once daily) but addition of boosting-dose ritonavir marks a new treatment.
	Follow-up	**Medium-term follow-up viral load**. The value closest to 168 days after the start of the new and continued treatment among those available within 112-224 days after the start of the new treatment.
		Medium-term follow-up CD4 counts. The value closest to 168 days after the start of the new and continued treatment among those available within 112-224 days after the start of the new treatment.

In addition to working on the Standard Datum table as agreed upon, project partners developing prediction models are free to use complete data from the integrated EuResist database available through the IBM Haifa Research Laboratories server. For example, this will let the partners use complete treatment history and past genotype data via patient-based related tables in order to train their model on an expanded set of "baseline" variables.

2.1. What to predict via training by the Classical Standard Datum:

- A ranking of the most active drug combinations (the end user is allowed to exclude any individual drug or directly select his/her favourite combinations). The ranking is based on expected success and failure based on the following definition: success is a drop of the viral load to undetectable levels (using the old threshold of 400 copies/ml) or a drop of more than 2 logs with respect to baseline value at short-term follow-up (8 weeks since the start of treatment); failure is any change of the viral load different from reaching undetectable levels or dropping of more than 2 logs with respect to baseline value at short-term follow-up (8 weeks since the start of treatment)
- The probability of achieving an undetectable viral load at 8 weeks with a measure of confidence if the end user provides the baseline viral load
- The expected level and change in viral load at 8 and 24 weeks with a measure of confidence if the end user provides the baseline viral load

- The probability of achieving a >1 log decrease in viral load at 8 weeks with a measure of confidence if the end user provides the baseline viral load
- The expected level and change in CD4 counts at 24 weeks with a measure of confidence if the end user provides the baseline CD4 counts
- The probability of achieving a >50% increase in CD4 counts at 24 weeks with a measure of confidence if the end user provides the baseline CD4 counts

3. The "alternative" Standard Datum

The "alternative" Standard Datum is based on the assumption that (i) any available genotype obtained while on therapy marks a failure of that therapy and (ii) any achievement of an undetectable viral load marks a success of the ongoing treatment. Note that the same treatment can first result in success (achievement of undetectable viral load) and then in failure (genotypic assay performed). This type of standard datum is expected to provide a large amount of records and is as follows (all data are mandatory):

Success (see Figure 2)	**New treatment** (first or change).
	Baseline genotype. The genotype closest to the start of the new treatment among those available within 90 days before the start of the new treatment.
	Time (days) to undetectable viral load while on the same treatment.

Failure (see Figure 2)	**New treatment** (first or change).
	On-treatment genotype. The first available genotype obtained after 28 days since the start of the new and continued treatment.

3.1. What to predict via training by the "alternative" Standard Datum:

- A ranking of the most active drug combinations (the end user is allowed to exclude any individual drug or directly select his/her favourite combinations)
- The probability of achieving an undetectable viral load at 8 weeks with a measure of confidence

Figure 1.

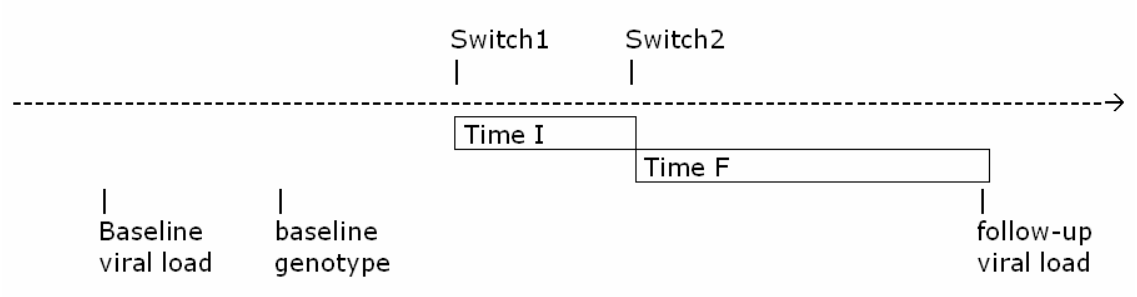

Figure 1. Graphic view of the requirements for baseline viral load and genotype in the Classical Standard Datum.

Two treatment lines have been used in the period between available candidate baseline and follow-up viral load. The treatment change event must be based around Switch2 (availability of follow-up viral load): time F must be 28-84 days and treatment #2 must be continuous. If time I is reasonably long (e. g. >14 days) it may make the use of baseline viral load incorrect with respect to Switch2. If time I is reasonably short (e. g. <=14 days) it can be safely assumed that treatment #1 did not significantly alter baseline viral load and baseline viral load is a valid baseline for the Switch2-centric treatment change event.

Figure 2.

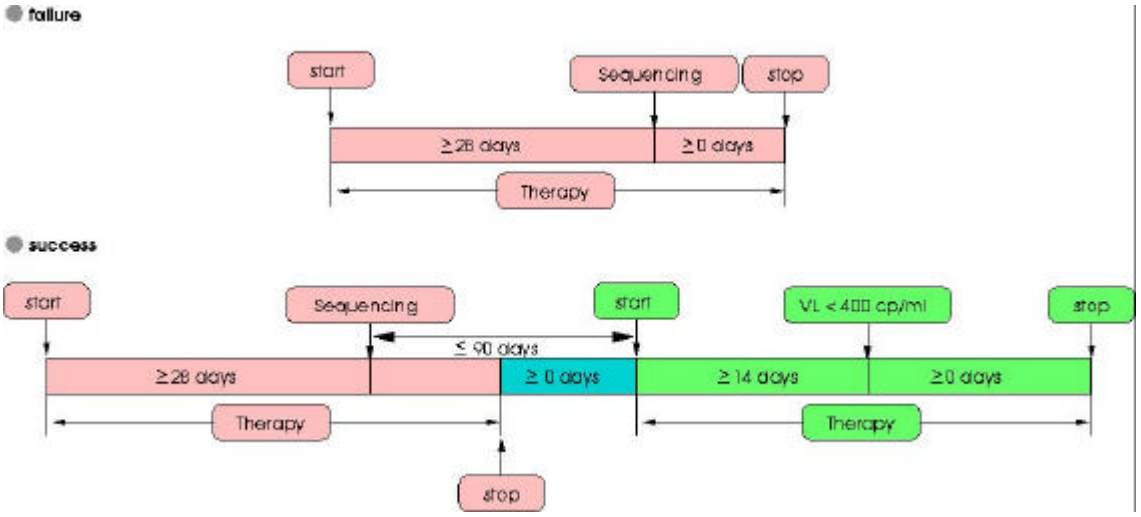

Figure 2. Definition of treatment success and failure in the Alternative Standard Datum.

Table 1.

Drug	Relative effectiveness
Zidovudine - AZT	1
Didanosine - ddI	1.1
Zalcitabine - ddC	0.3
Stavudine - d4T	1
Lamivudine - 3TC	1.4
Abacavir - ABC	1.4
Tenofovir - TDF	1.2
Emtricitabine - FTC	1.4
Nevirapine - NVP	1.4
Efavirenz - EFV	1.6
Saquinavir - SQV	1
Boosted saquinavir - SQV/r	1.6
Ritonavir - RTVI	1.3
Indinavir - IDV	1.3
Boosted indinavir - IDV/r	1.8
Nelfinavir - NFV	1.3
Amprenavir - APV	1.3
Boosted amprenavir - APV/r	1.8
Boosted fosAmprenavir - FPV/r	1.8
Atazanavir - ATV	1.3
Boosted atazanavir - ATV/r	1.8
Boosted lopinavir - LPV/r	1.9
Boosted tipranavir - TPV/r	1.9
Enfuvirtide - T-20	1.4

Table 1. Expected relative effectiveness of individual anti-HIV drugs approved for clinical use in Europe. Zidovudine (AZT) is arbitrarily set as the reference drug with value = 1.