The periodic report should cover the whole period since the start of the project, and clearly point out the progress achieved since the last report.

Total length: approximately 6 pages

Objectives of the report:

- Scientific: Demonstrate the scientific progress
- Administrative: Address administrative and /or management problems
- Financial: Show how the grant has been spent and justify the payments being made

1 General Information

File number:	047.016.009
Starting date of the project:	01.06.2004

Dutch project leader:	Prof. Cornelia M. van Duijn
Russian co-leader:	Prof. Tatiana I. Axenovich

2 Scientific Part

Overview of Research activities: Please specify which activities have been carried out.

(1) We have worked out a basic schema describing the software we aim to develop (June 2004). Two major blocks of the software are total-project data management block and single-user management / analysis blocks.
(2) We wrote a list of methodological, algorithmic and software problems we face in our research in Erasmus and Novosibirsk (August 2004 onwards). The list has been circulated between researchers from different countries, who work on similar topics, which allowed its extension (autumn 2004).
(3) One program for total-project data management (data rearrangement and verification) was developed by the August 2004. Later it was tested using real data from Erasmus and Novosibirsk projects. By now, another program is ready for use and several are at the testing stage.
(4) We have implemented a data generator (September 2004), which will allow simulation of the data obtained in a full-scale genetic research in a genetically isolated population. Using this data generator, we have tested performance of different database management systems (DBMS) for storage and retrieval of large amounts of genetic-epidemiologica data (December 2004). We have accessed the speed of data import and export and the amount of disk space required for storage using MySQL and MS SQL.
(5) In (4) it has been shown that using a standard DBMS is too expensive in sense that retrieval time was too long. Therefore a specialised binary data warehouse, which would exploit the nature of genetic-epidemiologic data under consideration was proposed. The performance of the prototype system is now being tested.
(6) Several freely distributed software packages were tested on Erasmus data (October-December 2004). These include programs for genotypic quality control and haplotyping.
(7) Two young Russian researchers have visited Erasmus for one month (August-September 2004). During this period they took courses in Genetic Epidemiology, met the members of Erasmus Lab and became familiar with the Erasmus research projects.

Scientific Results: What are the main results achieved and what is their scientific significance?

Include references to the list be1[1]low.

(1) For the purpose of distribution of software and knowledge developed during the project we created a preliminary version of the project's web page at http://mga.bionet.nsc.ru/NLRU/
(2) Several programs for the initial data management were created. These include pedigree structure verification and recoding program RECODE_PED (available from the Web). This program was tested using simulated and real data and is now distributed as a release candiadate. Other program distributed now as release candidate is POOL_STR (available from the Web), which allows pooling the data from several stages of genome scans performed using Short Tandem Repeats. Several progams are now at the beta stage: an interface for standard genotypic quality control program PEDCHECK, named GENOT_QC, and a program which splits pedigrees for their later use with Lander-Green (SPLIT_LG) and Elston-Stewart (SPLIT_ES) algorithms. A phenotypic quality control and description program PHENO_QC is currently in the initial development stage.
(3) Testing of free (MySQL) and commercial (MS SQL) DBMSs was performed on simulated data. We have accessed the speed of data import and export and the amount of disk space required for storage. It has been shown that these DBMSs are not suitable for managing large amounts of genetic-epidemiologic data. They need large disk space and a lot of time for importing information into database and handling queries. Report will be available on the Web within soon; we also started writing a manuscript on the topic.
(4) We have started developing our own binary data warehouse of genetic-epidemiologic data which will take into account specificity of data and configuration of queries.
(5) Standard freely distributed software packages were tested using Erasmus data. We have shown that genotypic quality control tool PEDCHECK can work well on our data. Other programs tested include MENDEL and PEDIG.
(6) A draft of list of methodological, algorithmic and software problems we face in our research in Erasmus and Novosibirsk was generated (available on the Web). The list has been circulated between researchers from different countries, who work on similar topics, which allowed its extension.

Publications: which scientific papers, presentations or patents have resulted directly from this project? Please note: papers which were published before the project started must not be included.

Journal publications:
Y.S. Aulchenko, A.M. Bertoli-Avella, C.M. van Duijn (2005) A method for poolin alleles frpm different genotyping experiments. Annals of Human Genetics (in press)
Web publications:
RN-list, a list of design, methodological and computational questions appearing in genetic-epidemiologic research in genetically isolated populations. http://mga.bionet.nsc.ru/NLRU/
Programs, published on the Web:
RECODE_PED, a program for verification of pedigree data and recoding pedigrees from free to standard format. http://mga.bionet.nsc.ru/NLRU/
POOL_STR, a program for pooling alleles from different genotyping experiments. http://mga.bionet.nsc.ru/NLRU/

1[1]

Please Summarise the scientific output (in numbers) below:

		Number of … with co-authorship	Number of … without co-authorship	**Academic Publications**
1	A	1		Publications in (international) refereed journals
	B			Publications in other (national) journals and other scientific output (abstracts in proceedings)
	C			Contribution to (chapters in) books
	D			Monograph
	E			Thesis (MSc, PhD)
				Professional Products
2	A			Patent
	B		2	Other professional products
				Other output
3	A	**1**		**Web publications**
				Conferences attended
4	A			