ORIGINAL ARTICLE

Annals of Agricultural and Environmental Medicine 2011, Vol 18, No 2, 207-214
www.aaem.pl

Application of advanced data collection and quality assurance methods in open prospective study – a case study of PONS project

Zbigniew M. Wawrzyniak[1,2], Daniel Paczesny[1], Marta Mańczuk[3], Witold A. Zatoński[3,4]

[1] *Institute of Electronics Systems, Warsaw University of Technology, Warsaw, Poland*
[2] *Faculty of Health Sciences, Medical University, Warsaw, Poland*
[3] *Department of Cancer Epidemiology and Prevention, Maria Skłodowska-Curie Cancer Centre and Institute of Oncology, Warsaw, Poland*
[4] *European Health Inequalities Observatory, Institute of Rural Health, Lublin, Poland*

Abstract

Introduction: Large-scale epidemiologic studies can assess health indicators differentiating social groups and important health outcomes of the incidence and mortality of cancer, cardiovascular disease, and others, to establish a solid knowledge base for the prevention management of premature morbidity and mortality causes. This study presents new advanced methods of data collection and data management systems with current data quality control and security to ensure high quality data assessment of health indicators in the large epidemiologic PONS study (The Polish-Norwegian Study).

Material and methods: The material for experiment is the data management design of the large-scale population study in Poland (PONS) and the managed processes are applied into establishing a high quality and solid knowledge.

Results: The functional requirements of the PONS study data collection, supported by the advanced IT web-based methods, resulted in medical data of a high quality, data security, with quality data assessment, control process and evolution monitoring are fulfilled and shared by the IT system. Data from disparate and deployed sources of information are integrated into databases via software interfaces, and archived by a multitask secure server.

Conclusions: The practical and implemented solution of modern advanced database technologies and remote software/hardware structure successfully supports the research of the big PONS study project. Development and implementation of follow-up control of the consistency and quality of data analysis and the processes of the PONS sub-databases have excellent measurement properties of data consistency of more than 99%. The project itself, by tailored hardware/software application, shows the positive impact of Quality Assurance (QA) on the quality of outcomes analysis results, effective data management within a shorter time. This efficiency ensures the quality of the epidemiological data and indicators of health by the elimination of common errors of research questionnaires and medical measurements.

Keywords

health indicators, epidemiologic study, information systems, data collection, quality assurance, health survey, Poland

INTRODUCTION

Large-scale population and epidemiologic studies can assess health indicators to define public health problems [1], and the ways in which lifestyles and living conditions determine health status need a comprehensive understanding [2-4]. This provides the basis for disease prevention activities, and understanding disease and health in the population [5, 6].

Health indicators may also include not only a measurements of illness or disease, but positive aspects of health (e.g. quality of life, life skills or health expectancy) and of health-related individual behaviors and actions, or social and economic conditions of the physical environment [7, 8].

Address for correspondence: Marta Mańczuk, Department of Cancer Epidemiology and Prevention, Maria Skłodowska-Curie Cancer Centre and Institute of Oncology, Roentgena 5, 02-781 Warsaw, Poland. Tel.: +48 22 546 26 16. Fax.: +48 22 643 92 34.
E-mail: manczukm@coi.waw.pl

Received: 16 October 2011; accepted: 05 December 2011

By using the outcomes from epidemiologic studies the differences in the health of populations can be observed. [9-11]. On the other hand, analysis of the important factors of these differences can increase achievable study objectives, and the extent of study objectives can be assessed from the outcomes of such a population study of health and disease [12-15].

The urgent need to understand the causes of these differences will allow the creation of a baseline for rational actions and health interventions, especially for the incidence and mortality of cancer [1,15-17], cardiovascular disease [6, 13], and other major causes of morbidity and mortality [14, 16, 18-20]. There is a high level of preventable premature morbidity and mortality in males, with marked differences observed between and within European Community countries, which can only be addressed by targeted activity across the lifespan [12].

The use of collected specific data is fundamental in epidemiology and population studies. Medical and social-related data are of different levels of abstraction and from

a wide range of sources (surveys, clinical trials, medical measurements, other medical checks and outcomes, or existing databases of any origin).

The lowest level of data abstraction in medical knowledge structure [21-23] (so called physical level) describes complex low-level data structures in detail. The next higher level of abstraction (the so-called logical level) describes relationships existing among those data in database in terms of a small number of relatively simple structures. The highest level of abstraction (view level) describes only a part of the entire database with remained complexity because of the variety of information stored in a large database [24]. Such structures of internal unknown relations are to be analyzed to increase knowledge about relations, and then to manage the association between health indicators and social-related issues [25]. Epidemiologic studies of causation use data in the search for information about the true nature of the relationship between exposure and disease, or influence of social issues on the incidence and mortality of some diseases.

The answer to the question of whether seen data is believable is determined by the critical question of how good is the quality of the data (outcomes). During a population study there are issues of population selection, measurement of study exposures, outcomes, or co-variates. Errors which may possibly occurr can lead to a biased estimate of the effect of exposure of risk for the disease of interest. Where association to be detected exists between all study participants, misclassification of exposure of a disease in outcomes can be randomly found. Finally, it can lead to an incorrect assessment of the relationship between exposure and disease.

The process of datasets production from an epidemiologic survey in the form of outcomes is data management, the methods of which are already widely known and available [26, 27]. Data management governs data-related processes during the survey and follow up research/analysis to set up the study, collect or enter data, and clean such data. The cleaning procedures process data until the study could be considered ready for analysis, and produces datasets with accurate data that reflect the values provided by the investigated sites. The principal responsibility of data management is to provide a clean database as a basis for statistical analysis, the relations between environmental aspects and socio-economic levels. The health outcomes described by some typical health indicators can then be assessed by using data analysis and data-mining techniques.

During the course of an epidemiologic study, it is a rule that an effort should be made to produce documents at key points during the study to record what was achieved, and to provide evidence of good practice. To ensure that the study document files at each of the data processes are consistent, Standard Operating Procedures (SOPs) are written that outline the contents of each study document file and database. SOPs ensure that studies (trials) are conducted, data generated, documented (recorded), and reported in compliance with the protocol, Good Clinical Practice (GCP), and the applicable regulatory requirements.

Quality assurance (QA) is the prevention, detection, and correction of data errors or other problems related to error reasons. QA, together with regulatory compliance for Medical Devices [28] are crucial, and a key requirement for quality methods is the creation of a data management plan (DMP). Additionally, a key requirement for Good Clinical Practice (GCP) [29] is the documentation of events during a study. The strict definition of QA taken from [29] is as follows: 'A planned and systematic process established to ensure that a study is performed and the data are collected, documented (recorded) and reported in compliance with this guidance and the applicable regulatory requirements'. In any population-based epidemiologic research the detailed DMP and appropriate procedures should be prepared and then run during its whole lifetime. The DMP and the required output documents are used as the starting point when conducting internal QA audits within the data management process. The automated QA software in the form of SOPs works especially during data collection and the aggregation/accumulation processes of databases to check data quality by the appropriate assessment measures (see the very useful handbook [30]).

The information technology (IT) infrastructure that is necessary to obtain outcomes of a population-based epidemiologic study needs to fulfill all the management needs, e.g. data collection, data transmission and storage, and processes of Data QA through electronic means.

In this paper are presented new advanced methods of data collection, management systems, and intensive data quality control and data security to ensure high quality data assessment of health-indicators in a large scale population study. The case study of the PONS project [31] is designed and prepared under the fulfillment of an entire IT and data management structure with central server and web-based peripheral applications, and tailored software managing data during entry, transmission, storage, and coding under the control of QA and security procedures.

MATERIAL AND METHODS

The material for the presented research is the project for a data-management structure design for a large-scale population study. On-going work is the investigation of using advanced databases technology with IT infrastructure for implementing an open-ended prospective study with very broad research branches in Poland (the PONS study [31]), and solving challenges found during the investigation. The design of the data management and management processes for the PONS project are experimentally applied in the project. The management structure has been constructed for such large-scale population study with its specific objectives [31] to successfully implement and establish a solid, high quality knowledge base for the prevention of major causes of premature morbidity and mortality.

The PONS project is aimed at a broad spectrum of various common chronic diseases by exposure and outcomes; the study data includes different personal data registered from self-administered questionnaires, a number of clinical measurements, and blood collected for storage in a modern Biobank to be use in prospective studies. From the other side, the IT system necessary for running the project can be remotely distributed by use of web-based mechanisms (client-server configuration). Therefore, it is proposed not only to manage data, but also data quality analyses within the entire IT infrastructure based on the database system. The implemented data quality by QA system includes data selection queries, joins, and aggregates, and QA programmes running in SPSS [31]. Through the QA procedures and the database IT structure, solutions can be achieved that provide the necessary connectivity between various distributed entry

data points and electronic patient health records; and enable the development and implementation of the knowledge base of health indicators.

The needs and requirements for IT system and data management applied for the project design are as follows: flexible and scalable IT system with remote acquisition from remote data collection points, supported by advanced web-based methods. IT system with a tailored design of software provides the promise for: 1) database aggregation of high quality medical data, 2) data security with quality control processes, 3) assessment of the data for the evolution of processes for monitoring the performance of data, outcomes and interviewers, and 4) assessment of knowledge quality projected in health indicators. In combination with proper organizational changes and skills, new data management design and services based on IT will become the key enablers of shared and continuous run of the project.

RESULTS

In this part, the most important layers of data-management structure implemented in the PONS study are described in detail. Our purpose is to show evidence of the advantages of such an approach to the data management by sophisticated and reliable solutions. The PONS study used electronic data collection for both the state of health questionnaire and medical measurements, forming an efficient type of database and computer as tools in data management.

Data flow in data management and QA

The proposed framework of the data management for PONS study is based on data transformation processes of a so-called raw data stream, collected from all information resources, to the final outcomes (information dataset) of the study governed by 2 types of quality standards. The first, with a general meaning, is the standard of good clinical practices, connected with quality standard for designing, conducting, recording and reporting trials that involve the participation of human subjects [29,32]. The second is the set of written quality control system protocols in the form of standard operating procedures. The overview of the context-aware applications is presented in Figure 1. This gives a visual representation of the functional relation of the system components when data are generated, documented (recorded) and reported during the PONS study.

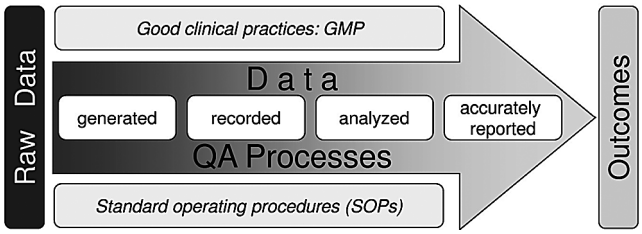

Figure 1. Data Flow environment for PONS study

The PONS study uses electronic data collection, data handling, and remote electronic trial data systems. Requirements for completeness, accuracy, reliability, and consistent intended performance (i.e. validation) of the data are ensured and documented by maintaining the standard procedures SOPs of these systems. Also, management of the electronic IT system maintains data security, preventing unauthorized access, an adequate data backup system, documents any necessary data changes due to QA cleaning without deletion of entered data. The documentation for conducting SOPs have been performed for system setup/installation (including the use of software, hardware, system operating manuals), system maintenance, data collection and handling with risk and quality assessments, validation and functionality testing, personnel training of computerized devices/systems used in the study.

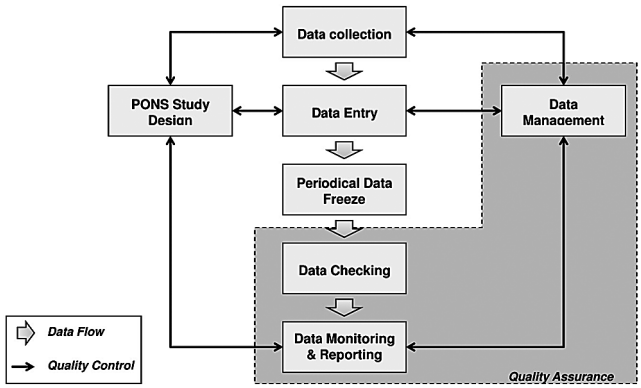

Figure 2. Data flow logics and quality assurance path

Quality control should be applied to each level and stage of data handling to ensure that all data are reliable and have been processed correctly and without errors (Figure 2). The processes of the data flow as collection, entry, temporal (periodical) freeze; checking and reporting/monitoring are under the QA standard procedures deployed as SPSS procedures and detailed protocols of data commutation. In our solution, the data management as a process itself is under the control of quality control with QA procedures, as well as the strategies and issues of the PONS objectives maintaining quality and security of data. To describe data management precisely, the dataset system and its logically and spatially heterogeneous sources is characterized in the paragraphs.

Models of datasets and databases

The principal responsibility of data management in the PONS is to provide a clean database; however, not only the quality of the database contents (i.e. study trial data and measurement outcomes), but also the database structure may affect the quality of the analysis. Data included in the study report must be traceable via the relational database and the case report form (CRF) to the source data at the investigator site or laboratory.

In the PONS, the data abstraction process, by which a data representation similar to its semantic meaning is undertaken, uses dataset structure of electronic participant record (EPR) (Figure 3), while hiding the implementation details characterized in the following paragraphs. The data at first layer in the EPR dataset in divided and stored in 2 parts; at the second layer, the health state questionnaire (HSQ) and Clinical Measurements, and finally divided into 3 datasets of HSQ, Medical Check-up and ECG Assessment (separated databases in our IT structure) from 2 providing servers: Questionnaire and SFTP server, and Medical Measurement Server (about 40 modules of KSSOMED system [33]). The logical architecture of EPR is presented in Figure 3.

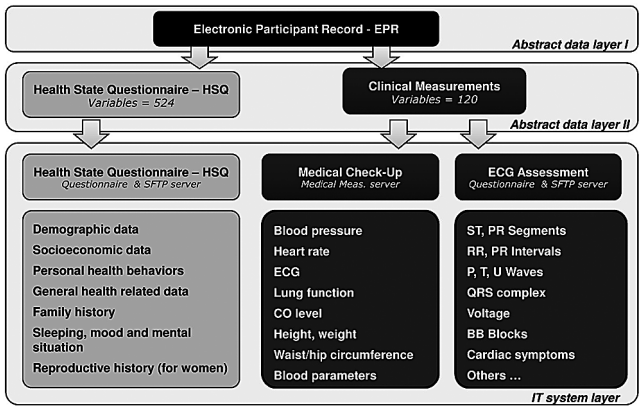

Figure 3. Data structure in EPR databases

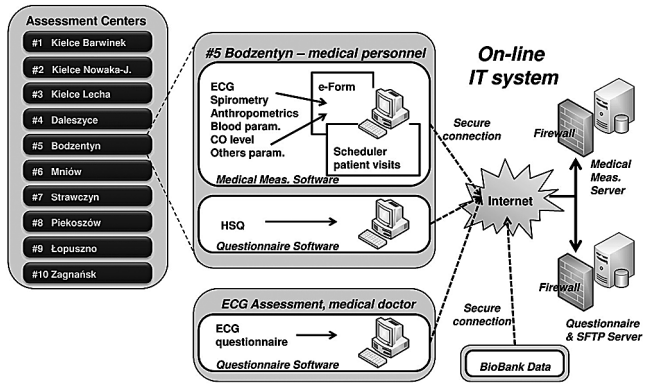

Figure 4. Structure of on-line IT system

IT on-line model

Data entry points of electronic data collection for both HSQ and Medical Measurements are situated in several facilities located in Świętokrzyski Province. In each of the 10 outpatient clinics, so-called Assessment Centres (e.g. in Bodzentyn) (Figure 4), the e-form consists of ECG, Spirometer checks, blood parameters, anthropometrics parameters, CO level, and other sample outcomes, and interviewer-administered on-line questionnaire are processed using a protocol of secure connection for the transmission of results to the central servers located in Warsaw. The other 2 components of data entry structure are: 1) ECG Assessment remote point with ECG questionnaire transmitted to the servers in the form of e-form assessment questionnaire, and 2) a Biobank in Kielce with its specially designed spreadsheets, which are stored in a central server in Warsaw using a secure Internet connection. Both of the above-mentioned computer systems work independently.

Data security

Security policy and procedures were successfully developed deploying applications across the entire IT infrastructure for database security. The specially designed database form is used to charge the central database. By these implemented solutions for collection, transmission, and data control in the PONS project, data security are as follows: access to the systems is authorized by login and password or personal identification number PIN. Additionally, in some critical points of the IT system, an IP address filtration is used. A secure encrypted https (SSL protocol) connection and VPN connection are used in communications between the servers and the clients (browsers and application for medical measurements collection), depending on the server applications used. Moreover, as a module of IT system secure policy, the *s*erver firewalls (in questionnaire server – 3 stages of firewall) and antivirus software on client computers are implemented. A secure FTP server is established for exchange and files storage, and additionally the secure data USB drive (256 bits hardware encryption) for files and data in manual exchange are applied. Both computer systems automatically back-up the databases (Figure 4) every day to prevent unexpected data loss.

Data collected from HSQ and ECG questionnaires by the open software survey tool Limesurvey [34] provide a variety of functionality developed to facilitate the design, administration, collection and management of the HSQ database. Central server databases manage the data by the web-based software, together with electronic data collection. This approach was useful for facilitating the complex skip patterns used in the PONS HSQ and ECG questionnaires, as well as in some built-in validity checks during data collection processes and Quality Assurance (QA) procedure. The HSQ and ECG questionnaire designs and functionality of both IT systems have been verified during the pretest, and are still run by QA monitoring.

Model of QA in data management

The quality assurance strategy in the PONS was developed based on determination of the data items causing problems, and is carried out by a full data audit from all the resources (PCs/server, KS-Somed, IT system software, HSQ and ECG questionnaires) and permanent inspection of every new data item (data, information, record, database, as well as paper record). Offending data is profiled by examination of the data diversity in a given data item, and determination of what the data should be by the use of acceptable domain of values according to a list of all acceptable values (without 'refused' and 'don't know'). Also, verification rules that constitute acceptable data were constructed, and rules for final cleaning determined how to finally clean the data. The final data after such procedures are used for preparation of cleaned databases as outcomes of the PONS study.

Quality Assurance processes performed in the Pre-Data Collection phase have required well-established quality control procedure on the following: 1) design and review of the specific HSG and ECG questionnaires (Figure 3), 2) programming of the questionnaires into the IT system (deployment) and 3) sample design and pretest. The second phase of Data Collection and Management has been operating until now (the project is still on-going) and consists of the following procedures:

1. setup and maintenance of the IT management system and questionnaire software;
2. conduction of the interviews and data upload to the IT systems;
3. aggregation and preparation of the raw data file for cleaning;
4. data cleaning built-in procedures.

The Post-Data Collection phase is designed to use all the IT and QA resources to process:

1. procedures for cleaning and preparing the data for sample weight calculations;
2. assessing the quality of the sampling, sample weights and assessment of sampling, non-response and non-sampling errors;

3. creation of the final analytic data file consisting of the cleaned outcomes.

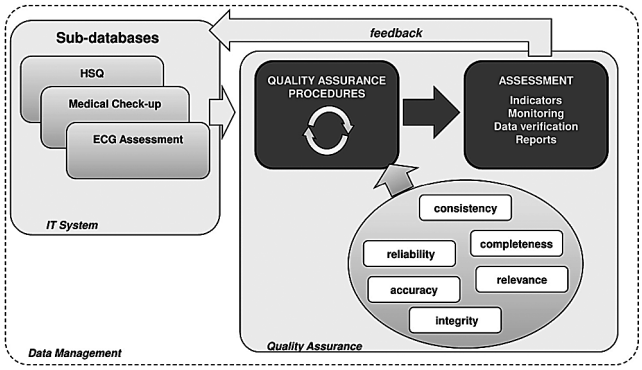

Figure 5. Data management and quality assurance

During the current phase of the study lifecycle, the logical structure and content of databases were analyzed by QA procedures for data profiling. The overview of the context-aware application is presented in Figure 5, and gives a visual representation of the QA functional relation of the data management system (cf. Figure 1 for relation of QA with data flow management for the study lifecycle). The well-known idea of information feedback from QA procedures means profiling source data, and provides extensive capabilities of column analysis (variables value), row analysis (patient record), primary key analysis (personal identification number PESEL as a unique identifier of the record and relational access to HSQ, ECG, and Biobank datasets), and cross-domain analysis (examines content and relationships across HSQ, ECG, and Biobank datasets). The above-mentioned types of analysis generate a full report, and monitoring indicators DVR (data verification report) based on frequency distributions and examination of all values to infer its definition and properties, such as domain values, statistical measures, and minimum/maximum values. Each variable column of every dataset is examined in detail. The following properties are observed and recorded as measures or indicators: basic data types, count of distinct values or cardinality, count of empty, null/non-null values, minimum/maximum, and average/median numeric values, and length measures for nominal and text data. Domain analysis characteristics determine the data domain values for any data element and their aggregate measures as the frequency distribution, number of occurrences, and measures of correspondence to a value in a dataset to find anomalies in datasets. With the designed data management structure and current outcomes of verification, validation rules for data and evaluate data sets for compliance are created. These rules assist processing operations for monitoring purposes (not only for QA) in creating metrics run and track over time for defined relationships between and among data (including metrics for interviewers). Beyond the requirement of data (see characteristics in insert in Figure 5), validation rules in the PONS in data profiling can check the data conforms to certain constraints as containment of certain value/string, equality to a certain value, existence/occurrence, data range, integrity reference and reference uniqueness.

Entire IT structure model integrated with QA

The architecture of the entire IT system and data functionalities of the PONS is presented in Figure 6.

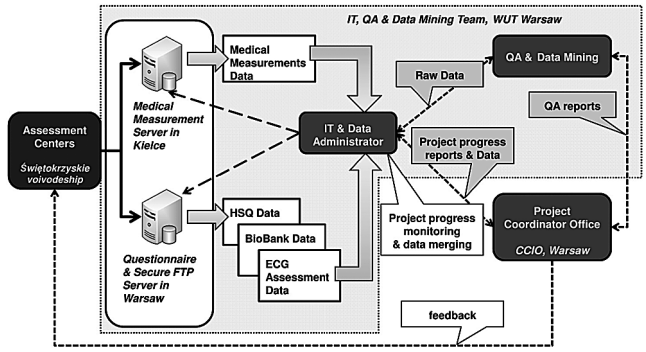

Figure 6. Data management and quality assurance in entire IT and organizational structure

As a final design of the IT remote data-distributed system, the Figure presents 4 source databases of Medical Measurement Data, HSQ Data, ECG Assessment data, and Biobank data from 2 data providers: servers (Medical Measurement Server in Kielce, and Questionnaire and secured FTP server in Warsaw). Each database stores data on servers (Figure 4) located inside the IT system architecture at Kielce and Warsaw, under the control of an IT & Data Administrator. The portion of new raw data for QA & Data Mining, on-line accessibly on the central server, are frozen at every data extent for a patient number (patient ID) of about 1,000. Results of QA procedures are run on a separate computer system and thus securely provided to the Project Coordinator Office for the monitoring and analysis of further outcomes resulting in feedback to Assessment Point of IT and QA Team. Therefore, each database is autonomous, and the level of data management of the Project Coordinator Office controls access to data and any cleaning procedures. Logical algorithm in detail of QA of processes' management and outcomes is shown in Figure 7. New data entered for the IT system by medical personnel should be aggregated to a appropriate database (indexed by an IP number, such as PESEL), but with a status which could be named, as much as possible, as very acceptable due to the quality, completeness and integrity at the accepted level of data quality described by statistical measures. The latest captured archive revisions (data freeze in Figure 7) on a certain portion of the data requested by QA & Data Mining Team allow the use of a common set of data for QA calculations and analyses. From the assessment of the QA measures, a decision about the level of intervention/control and its subject (on medical personnel, or only on new data portion) is undertaken depending on the state of alarm or warning.

The proper use of SOP is described in the management manual of QA policy. Final findings of last data freeze associated with publication of study outcomes allows publication of data cleaned at the Post-Data Collection Phase.

In content aspects of the PONS study, outcomes of cross-tables and event indicators in QA are practically applied in identifying key process variables for the data collection and processing phase. The main indicators are percentage of detected/corrected errors, overall coding accuracy rate, counts of data without 'missing', 'within the range' items, and data without 'don't know' or 'refused' responses. Essential factors of the quality control in the PONS study are the following: each staff person is familiar with all procedures by specially designed 'question by question' instructions for

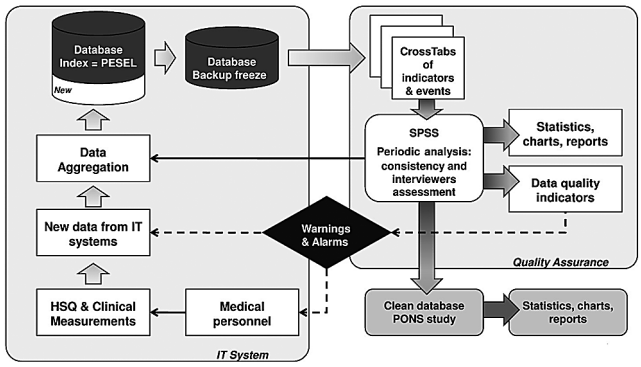

Figure 7. Quality assurance processes in IT and study

interviewers, and written manuals of operations for databases and detailed descriptions of the procedures (SOPs) to each data collection instrument and its databases. The number of created user accounts in the IT collection system is about 40. For the question path-flow in the questionnaires, the presence of sensitive questions and filtering questions directs the flow. The specially deployed procedures of the permanent QA monitoring, preceded by initially performed pretest, assess the flow of processes, and appropriateness of variable categorizations in the questionnaires by a content check.

During data collection in the PONS project management, the deliverables in the form of different reports and other descriptive documents are produced, as well as some server upgrades or special-purposes software block for the overall project are built, namely:

1. raw data and QA codebooks;
2. finally recoded QA reports;
3. non-valid items reports;
4. short and regular QA analysis reports;
5. descriptive and percentile statistics reports,
6. Data Verification Form (DVF), for Medical Check-Up outcomes;
7. additional requests for information reports (for HSQ outcomes);
8. short and regular progress reports;
9. interviewer assessment reports;
10. improvement recommendation reports.

Progress of the PONS as of 25.09.2011 is the record status of 8,650 patients examined, and the entire IT system has had 380 days of continuous and stable computer system running without one non-operation day.

Implementation of follow-up control with data freeze backup can assess the data quality on the PONS sub-databases at the level of correctness of 99.8%(for Medical check-up when n= 6,925), 99.8% (for HSQ when n=7,606) and 99.95% (ECG Assessment when n=6,568). The results contain data of the last available QA report release on 25.09.2011.

DISCUSSION

The most effective data collection process starts with a good experimental framework design of data and study management. Our experience with HEM, GATS, ECAP and other epidemiologic and population projects has led to the following recommendation: for a research community, good data quality translates into a solid knowledge structure of high quality for the health interventions, and especially for the prevention of major causes of disease morbidity and mortality.

A quality management system applied to the data flow is the procedure and process necessary to implement, assess and ensure data quality. However, the impact of the structure of survey questions (survey path in questionnaires and its relation with the medical measurement outcomes) requires some specific solutions in the design of the computer IT system framework. The functional needs of the PONS content and technical requirements of the PONS study data collection and data quality are fulfilled and shared by proper design and analytical system solutions, as described above, to obtain data reasonableness and data consistency. The IT systems are deployed to be secured at each level against unauthorized access from third parties, and data loss prevention procedures monitor the evolution of the databases.

The integration of medical data structures for an abstract level of sources could be an advantage for such an approach of the entire IT framework structure integrated with QA. Data are taken in the well-known form of Electronic Health Records EHR [35] as electronic participant records (EPR). The concept of EHR is a longitudinal record of patient health information (generated at any data entry point) and of population parameters. It automates and streamlines the clinician's workflow, and supports other care-related activities, including evidence-based decision support, quality management, and outcomes reporting. From information workflow, the EHR record is embedded in the network-connected enterprise-wide information systems. In the presented case, it also consists of medical check-up outcomes for the purpose of assessment of health indicators related to environmental aspects and socio-economical levels. Logical data integration of different origin and the fusion of collected and measured data in the form of relation databases are specially designed to manage the processes of medical data knowledge building. Thus, the findings from the data mining processes can describe the relations between parameters of disease, symptoms, and health status in a population by some health indicators and the health outcomes.

This systematic process approach developed for the design of the data management framework in the PONS study is useful to successfully facilitate the complex skip patterns used (especially in the HSQ and ECG questionnaires), as well as in some built-in automatic validity checks during data collection and QA procedures. The reporting system of wide-range measures has been developed to monitor the progress of the PONS study. The IT systems are secured at each level against unauthorized access, while the data loss prevention procedure monitors evolution of the databases.

The above-mentioned functionalities of survey software (both system and device), together with data verification procedures, assure a high level of quality during data collection and data aggregation with the above- mentioned data freeze backup process under QA control.

By providing a framework around which all IT processes can be standardized and controlled, a design structure has been developed that enables the PONS study to embark on new healthcare initiatives with outcomes of higher data quality. Data monitoring and analysis ensure that the whole of the IT infrastructure and processes are functioning within known parameters so that study levels are maintained or errors corrected in many cases before the situation becomes critical. Data quality control processes and software enhancements

and developments can be linked to quantitative data and outcome analysis rather than subjective feedback.

The computer system was designed and built within the project framework requirements and the epidemiologic assumptions for such a population study. The adopted solutions are presented in detail below in the Results section. The most important data management tasks undertaken by the entire computer IT system can be summarized as follows:

1. effective screen interface data forms to enter the data from the survey questionnaires and medical check-ups;
2. a suitably efficient concept of scalability to increase the total throughput under an increased data load when assessment centres are added, and to manage visit schedule by a graphic tool;
3. automatic data validation at the load to the system;
4. implementation of effective methods of reporting and verification of data quality;
5. security of the system during transmission and access at every stage of processing.

These issues have been successfully implemented throughout our entire IT computer system and have been running very smoothly. Within the given time constraints accompanying the system building, the majority of proposed software solutions were selected from known applications, but necessarily adopted and properly parameterized for the project management requirements. Some elements of the system were constructed specially for this project, primarily connected with run and verification rules for question path-flow in the questionnaires and QA rules. An important testing issue of each computer system running in the population study is a pretest to evaluate the functions of the system, efficiency of individual components, and the correctness of data entries and database structures. The more efforts that are devoted to the examination and testing of disclosure errors in the pretest survey outcomes, the better the efficiency achieved in the main study. For preliminary examination, preparation and check of the validation rules of data quality and data security should also be considered. Nevertheless, examination of data quality and data and entire system security protection is a continuous process that should be maintained throughout the whole time of the study.

Despite the assumed correctness of the system solutions and fulfillment of developed requirements, the system can recognize that some elements and issues could be improved by permanent monitoring, this would bring result in a greater effectiveness of the IT and QA team. As the entire IT system consists of several separate elements and procedures acting independently, the data entry at the time of collection and aggregation within the system are stored in different databases. The ongoing reporting in the monitoring procedure checks the data consistency across different databases. This functionality can be improved in the central database by entering the automatic synchronization mechanism of autonomous databases when downloading data from collection points (assessment centres). The functional extent of this mechanism would be automatically detect discontinuities and inconsistencies in aggregated data. Presented functionality would simplify and speed-up the process of checking data integrity and consistency, which obviously is part of the post-collected procedure by QA methods. Implementation of such a mechanism in the form of automatic checking rules should be deployed in the system to decrease the possibility of potential errors and improved performance of individual part in data processing.

The suggested model addresses the need for tight control over data flow through a minimal increase in management tasks. This is based on structured data collection forms, a theoretical data flow, and a modular combination of medical measurement and patient administrative data management, and provides a framework within which the statistical stability of interim reports can be documented.

CONCLUSIONS

This article presents a new advanced method of practical design for the management of data collection and a management system with intensive data quality control and data security to ensure high quality-data assessment of health-indicators in large-scale population study. The case study of the PONS project is designed and prepared to fulfill the requirement of data management and IT structure with central server and some web-based peripheral applications, and tailored software managing data during entry, transmission, storage, coding under the control of Quality Assurance with data security procedures. In the proposed solution for project management and data collection, the aspect of data security has been analyzed in detail, and some advanced in IT system secure solutions have been implemented.

The use of modern advanced database technologies and remote hardware/software structure efficiently and successfully supports the research of large-scale epidemiologic surveys. The collection of medical data, with related environmental and socio-economical information from different data sources, results in medical a data knowledge structure especially useful for future health prevention and prognosis of critical health indicators in a population.

The data from disparate and deployed sources of information are integrated into databases of knowledge's structure via software interfaces, and archived by a multi-task server, while the designed processes (computer applications) supervise and carry out data security with quality assessment.

Development and implementation of follow-up control of data consistency and quality data analysis on the PONS sub-databases have excellent measurement properties, without error, at the level of 99%.

The practical solutions implemented in the entire IT system by custom software application show a positive impact of QA on effective data management, with shorter time, efficiency, and ensuring the quality of epidemiologic data and indicators of health with the elimination of common errors in research questionnaires and medical measurements.

Development and practical deployment of the PONS concept as a client-server web web-based IT data system, assisting and performing the research in accordance with the methodology of the epidemiologic surveys, have evidently been carried successfully, and have practically controlled the data quality problems.

The requirements of the above-mentioned epidemiologic surveys are be achieved through e-Health Information Technologies that play a decisive role in promoting the concept of effective data management in large-scale epidemiologic studies, providing the necessary connectivity between variously distributed points of data entry, and enable

the development and implementation of electronic personal health record systems.

ACKNOWLEDGEMENTS

The study was supported by a grant from the Polish-Norwegian Research Fund (PNRF-228-AI-1/07). The authors express their thanks to the members of the PONS project team and the participants for their contribution to the study.

REFERENCES

1. Powles JW, Zatonski W, Vander HS, Ezzati M. The contribution of leading diseases and risk factors to excess losses of healthy life in Eastern Europe: burden of disease study. BMC Public Health 2005;5:116.
2. Zatonski W. The East-West Health Gap in Europe--what are the causes? Eur J Public Health 2007;17(2):121.
3. World Health Organization, Health Promotion Glossary, WHO, Geneva, 1998, available from: http://www.who.int/hpr/NPH/docs/hp_glossary_en.pdf
4. Smith BJ, Tang KC, Nutbeam D, WHO Health Promotion Glossary: new terms, Health Promotion International, 2006;21(4):340-345, doi:10.1093/heapro/dal033.
5. Zatoński W, Didkowska J. Closing the gap: Cancer in Central and Eastern Europe (CEE). Eur J Cancer 2008; 44:1425-1437.
6. Zatonski W, Campos H, Willett W. Rapid declines in coronary heart disease mortality in Eastern Europe are associated with increased consumption of oils rich in alpha-linolenic acid. Eur J Epidemiol 2008; 23(1):3-10.
7. Bambra C, Eikemo TA. Welfare state regimes, unemployment and health: a comparative study of the relationship between unemployment and self-reported health in 23 European countries. J Epidemiol Community Health. 2009;63(2):92-8. Epub 2008 Oct 17, PMID: 18930981.
8. Arber S. Social class, non-employment, and chronic illness: continuing the inequalities in health debate. Br Med J (Clin Res Ed). 1987;294(6579):1069-73, PMID: 3107698.
9. Puig-Barrachina V, Malmusi D, Marténez JM, Benach J. Monitoring social determinants of health inequalities: the impact of unemployment among vulnerable groups. Int J Health Serv 2011;41(3):459-82, PMID: 21842573.
10. Caiazzo A, Cardano M, Cois E, Costa G, Marinacci C, Spadea T, Vannoni F, Venturini L, [Inequalities in health in Italy]. Epidemiol Prev 2004;28(3Suppl):i-ix,1-161. PMID:15537046.
11. Zatonski W, et al. (Eds.). Closing the health gap in European Union. Cancer Center and Institute of Oncology, Warsaw; 2008. www.hem.waw.pl
12. White A, de Sousa B, de Visser R, Hogston R, Aare S, Makara P, et al. (Eds). The State of Men's Health in Europe, Luxembourg, The European Commission 2011 http://ec.europa.eu/health/population_groups/docs/men_health_report_en.pdf.
13. Zatonski WA, Willett W. Changes in dietary fat and declining coronary heart disease in Poland: population based study. BMJ 2005;331:187-8.
14. Zatonski W, Mikucka M, La Vecchia C, Boyle P. Infant mortality in Central Europe: effects of transition. Gac Sanit 2006;20:63-6.
15. Zatonski WA, Manczuk M, Powles J, Negri E. Convergence of male and female lung cancer mortality at younger ages in the European Union and Russia. Eur J Public Health 2007; 17(5):450-4.
16. Zatoński W, (Eds.) with: Mańczuk M, Sulkowska U, and the HEM Project team. Closing the health gap in European Union, Cancer Center and Institute of Oncology, Warsaw 2008.
17. Didkowska J, Manczuk M, McNeill A, Powles J, Zatonski W. Lung cancer mortality at ages 35-54 in the European Union: ecological study of evolving tobacco epidemics. BMJ 2005;331:189-91.
18. Bosetti C, Levi F, Lucchini F, Zatonski WA, Negri E, La Vecchia C. Worldwide mortality from cirrhosis: An update to 2002. J Hepatol 2007;46:827-39.
19. Jha P, Peto R, Zatonski W, Boreham J, Jarvis MJ, Lopez AD. Social inequalities in male mortality, and in male mortality from smoking: indirect estimation from national death rates in England and Wales, Poland, and North America. Lancet 2006;368(9533): 367-70.
20. Zatoński WA, Sulkowska U, Mańczuk M, Rehm J, Boffetta P, Lowenfels AB, La Vecchia C. Liver Cirrhosis Mortality in Europe, with Special Attention to Central and Eastern Europe. Eur Addict Res 2010;16:193-201.
21. Horn W, Buchstaller W, Trappl R. Knowledge structure definition for an expert system in primary medical care, Proceeding IJCAI'81 Proceedings of the 7th international joint conference on Artificial intelligence, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA 1981;2:850-852.
22. Swe T, Swea M, Sai N, Kham M. Case-based medical diagnostic knowledge structure using ontology, The 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010:729 – 733.
23. Lavrac N, Keravnou E, Zupan B. Intelligent Data Analysis in Medicine, 2000, 1-62, available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.175
24. Wojtyła A, Biliński P, Jaworska-Łuczak B. Regulatory strategies to ensure food and feed safety in Poland – update review. Ann Agric Environ Med 2010;17:215–220.
25. Raphael D. Shaping public policy and population health in the United States: why is the public health community missing in action? Int J Health Serv. 2008;38(1):63-94, PMID: 18341123.
26. Prokscha S. Pract ical Guide to CLINICAL DATA MANAGEMENT, CRC Press, Boca Raton, 2007.
27. Garcia-Molina H, Ullman JD, Widom J, Database Systems: The Complete Book, 2/E: Prentice Hall, 2009.
28. U.S. Department of Health and Human Services, Food and Drug Administration, Guidance for Industry - Computerized Systems Used in Clinical Investigations, May 2007, available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070266.pdf.
29. FDA, Good Clinical Practice VICH GL9, Guidance for Industry GOOD CLINICAL PRACTICE VICH GL9, June 2000, available from: http://www.fda.gov/downloads/AnimalVeterinary/GuidanceComplianceEnforcement/GuidanceforIndustry/UCM052417.pdf
30. United States Environmental Protection Agency, Data Quality Assessment: Statistical Methods for Practitioners EPA QA/G-9S, Washington, DC 20460, EPA/240/B-06/003, February 2006, available from: http://www.epa.gov/QUALITY/qs-docs/g9s-final.pdf
31. Project PONS, official web-site, available from: http://www.projectpons.pl/en_pages.html,3,1,about-the-pons-project.
32. European Medicines Agency, Guideline for Good Clinical Practice, ICH Topic E 6 (R1), July 2002, available from: http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002874.pdf
33. Integrated computer system of clinic management KS-Somed, KAMSOFT SA, available from: http://www.kamsoft.pl/prod/somed/info.htm
34. LimeSurvey project, LimeSurvey 1.91+ release, http://www.limesurvey.org
35. Gunter TD, Terry NP. The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions, J Med Internet Res 2005;7(1): Published online 2005 March 14. doi: 10.2196/jmir.7.1.e3, available online from: http://www.jmir.org/2005/1/e3/.