Registries in Clinical Epidemiology: the European Surveillance System on Contact Allergies (ESSCA)

W. Uter1; A. Schnuch2; M. Wilkinson3; A. Dugonik4; B. Dugonik5; T. Ganslandt6

1Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany; 2Information Network of Departments of Dermatology, Georg-August University of Göttingen, Göttingen, Germany; 3Department of Dermatology, Leeds Teaching Hospitals NHS Trust, Leeds, UK; 4Department of Dermatology, University Medical Centre Maribor, Maribor, Slovenia; 5Faculty of Electrical Engineering and Computer Science at the University of Maribor, University of Maribor, Maribor, Slovenia; 6Medical Center for Information & Communication Technology, Erlangen University Hospital, Erlangen, Germany

Keywords
Registry, dermatitis, allergic contact dermatitis, epidemiologic surveillance

Summary
Background: Disease registries rely on consistent electronic data capturing (EDC) pertinent to their objectives; either by using existing electronic data as far as available, or by implementing specific software solutions.

Objectives: To describe the current practice of an international disease registry (European Surveillance System on Contact Allergies, ESSCA, www.essca-dc.org) against different state of the art approaches for EDC.

Methods: Since 2002, ESSCA is collecting data, currently from 53 departments in 12 countries. Departmental EDC software ranges from spreadsheets to comprehensive “patch test software” based on a relational database. In the Erlangen data centre, such diverse data is imported, converted to a common format, quality checked and pooled for scientific analyses.

Results: Feed-back to participating departments for quality control is provided by standardised reports. Varying author teams publish scientific analyses addressing the objective of contact allergy surveillance.

Conclusions: Although ESSCA represents a historically grown, heterogeneous network and not one unified approach to EDC, some of its features have contributed to its viability in the last 12 years and may be useful to consider for similar investigator-initiated networks.

Correspondence to:
Wolfgang Uter
Department of Medical Informatics, Biometry and Epidemiology
Friedrich-Alexander University Erlangen-Nürnberg
Waldstr. 4 – 6
91054 Erlangen
Germany
E-mail: wolfgang.utter@fau.de

1. Introduction

Disease registries used as a basis for scientific analyses rely on consistently collected data pertinent to their objectives. The extraction of relevant information from electronic patient record systems in terms of “routine data” for purposes of a registry, or also for an ad-hoc clinical study, is an important field of medical information science research and service [1]. However, if the scope of such available data does not fully suffice to meet the respective research needs, study- or registry-specific data has to be collected in addition, and merged to the existing data. For some research topics, virtually no “routine” data may be available beyond basic patient characteristics such as age and sex. This was the case almost 30 years ago, and is still the case nowadays, according to the best of our knowledge, concerning data used for the clinical surveillance of contact allergy [2]. Such data is generated during the diagnostic work-up of patients with suspected allergic contact dermatitis by means of the so-called patch test [3].

The following article addresses the evolution and the current practice of the European contact allergy surveillance network ESSCA (www.essca-dc.org). The technical details (standard operating procedures, structure of central data pool) are presented briefly, but shown in detail in a web document (www.essca-dc.org/doc/essca_central_SOPs_online.pdf, last accessed 2016-01-27). The ESSCA data centre may serve as an example of the technical difficulties, and possible solutions, which are encountered when trying to collect consistent clinical data from a heterogeneous multinational environment, specifically, in the absence of funding – beyond an initial start-up period funded by the EU (QLK4-CT-2001-00343) 2002–2004 – which would otherwise enable implementation of a more uniform and efficient structure along the lines of concepts discussed below. Some characteristics of ESSCA will be discussed which are believed to contribute to sustainability and notable scientific output even from such a
non-supportive background and may be of value also for other research networks.

2. Background

2.1 General Options for Setting Up a Disease Registry

Electronic data capture (EDC) solutions in a registry network can be categorised according to the following two dimensions:

- Data hosting: Locally installed (uniform or diverse) software, potentially incurring significant manual efforts to export, transfer and harmonise datasets for central pooling and maintaining multi-site software installations versus one centralized EDC solution collecting data in a central data repository. Many successful implementations of such EDC platforms in the context of disease registries have been reported [4–6].

- Level of integration into departmental routines: A “minimal add-on” strategy minimising the amount of data entered additionally for the research purpose, e.g. by collecting only aggregated or a priori selected patient information, with potential limitations concerning the future use of data versus a “replacement” strategy, by which the research EDC instrument shall cover all departmental documentation needs, thus eliminating the necessity for additional (paper) routine documentation [7, 8].

However, regardless of whether locally installed software or a centralized EDC platform is used, these approaches still require data entry into a dedicated research platform, which may duplicate documentation requirements of the actual patient care work-flow. The “replacement” strategy described above aims towards using the dedicated research platform to also cover local departmental needs, but it may be difficult to cover all such needs encountered in a multinational research network in one single software solution. Some requirements may be impossible to address (e.g., legally mandated documentation in local billing or quality control systems); implementation of interfaces such as HL7 may provide some limited integration in this situation.

Concerning research topics other than contact allergy where this is feasible, a “secondary use” strategy could be implemented, which relies on data which is already routinely documented in the electronic health record (EHR), made available for research use. The use of routine data could be regarded as following a “minimal add-on” strategy, in terms of adding no additional burden of documentation on the local staff. Data collected in such a way can either be made available through exports and pooling as described above, or through federated approaches e.g. using the Shared Health Research Network (SHRINE) platform [9], in which data resides locally and is queried remotely. Applicability of this approach, however, relies on the overlap between routine documentation and research needs, which depends on both research focus and routine documentation requirements. Also, the re-use of existing documentation items implies that data elements and value sets have to be taken “as is” and cannot easily be harmonized to better meet research requirements.

A more comprehensive approach would be to implement a “single source” documentation strategy, in which all data elements required both for routine care as well as associated research projects are entered in the EHR, and research data is consecutively extracted [10, 11]. As such, the “single source” strategy would mostly be a “replacement” (of a previous, less comprehensive and flexible EHR) strategy, while trying to minimise documentation efforts by providing a single “all-purpose” EHR, without, however, restricting the scope of data a priori as a “minimal add-on” strategy necessarily would. The advantage of this approach is that documentation can be optimized to meet the needs of both routine care as well as research, and that those data elements and value sets can potentially be harmonized across sites. Also, federated approaches like SHRINE [9] can be used in a single source context. However, single source approaches may be the most difficult to put into practice, requiring an EHR platform that can be adapted to local and dynamically changing documentation needs as well as support from local routine IT departments and/or vendors to implement the respective data entry forms and work-flows. Harmonizing this integration of clinical care and research documentation across multiple sites creates additional challenges both concerning differences in local work-flows and semantic issues [12].

2.2 Definition of Contact Allergy and Its Clinical Surveillance

Contact allergy is an acquired immunological alteration resulting in allergic reactions of the skin (allergic contact dermatitis) after contact with a sufficient amount of the “allergen” in those who became sensitised (allergic) during an earlier exposure to the specific trigger. Allergens comprise natural as well as man-made substances and abound in the (work) environment [13]. However, those affected by allergic contact dermatitis only partly consult the medical system [14]. Affected persons surfacing as patients are either treated (without further specialised and specific diagnostics) at the level of primary care, or are referred for diagnosis and treatment to a dermatologist, who may also be directly consulted, depending on the health system. The availability of services for cutaneous allergy will depend on the wealth and degree of development of a country and society. Normally dermatologist will perform a so-called patch test to identify the allergen, or sometimes the multiple allergens, causing the presumptive allergic contact dermatitis. At this particular point of medical care clinical and demographic data are collected which are relevant for contact allergy surveillance. A necessary set of variables has been defined as ‘minimal dataset’ by the European network ESSCA (www.essca-dc.org), consensualised at a general assembly in 2003, described in Table 1.

Morbidity data are collected in patients, and not in population samples. This implies, as in other clinical surveillance systems, that the relative frequency of diagnosed contact allergies cannot be interpreted in terms of prevalence at the population level. Indeed, prevalence is expected to be higher to a varying degree, due to morbidity-driven patient selection [15] which may also vary between countries and departments, respectively [16]. The resulting considerations involved in the use
and interpretation of the clinical epidemiology of contact allergy are beyond the scope of the present paper and have been discussed elsewhere [2, 17]. Some examples of contact allergy surveillance will be briefly presented in the results section.

3. Methods

3.1 Electronic Data Capturing Software for Contact Allergy

As mentioned, only a minute part of the ‘minimal dataset’ for contact allergy is potentially covered by data collected routinely, as we are aware of. This covers mostly just basic demographic and clinical data, and possibly the patch test series applied – however, no detailed results. Hence, all local EDC software solutions have specifically been developed for the purpose of documenting patch test results and the associated demographic and other clinical data (Table 1). At present, both the “minimal add-on” and the “replacement” strategies are being followed, depending on the preferences and scope of the respective research networks. The varying degree of flexibility (and complexity) of the “patch test software” used and the consequences for data analysed is described in [18]. Briefly, the following options are being used by contributors of the ESSCA network, ordered by increasing complexity (and: flexibility, usefulness and need for IT support):

- Use of an electronic spreadsheet, following a “minimal add-on” strategy. In this spreadsheet the rows usually represent single consultations of patients. The columns list different attributes of this patient, such as age, gender, occupation, and patch test reactions. Only one or, at maximum, two patch test reading results per allergen are usually included, covering just the baseline series applied routinely to every patient, but no specialised test series.

- A relational database used as a back-end, with workflow-oriented screen forms as a front end, programmed e.g. with MS-Access, as used by the British Society of Cutaneous Allergy (http://www.cutaneousallergy.org/, last accessed 2015–06–09). This potentially more flexible way to collect patch test data allows for the inclusion of (different versions of) several test series applied to a patient. Some standardised reports of the locally collected data are available (e.g., results with a certain test series in a defined period). Data can be exported for central pooling, covering defined periods.

- As another example of a software based on a relational database, the multilingual WinAlldat software [19] is used by several departments of the European Surveillance System on Contact Allergies (ESSCA; www.essca-dc.org) network, enabled by start-up funding by the EU (QLK4-CT-2001-00343; 2002–2004). The software allows administration of centre-specific test series and individualised series with patients’ own materials, linking between different consultations of one patient (n:1), network integration, an interface to hospital information systems and, most recently, a data mining tool to examine the contents of the database with a set of freely definable filters (“replacement” strategy). Anonymised export into csv files is provided. However, “WinAlldat/ESSCA” is locally installed, which implies performing any updates in each participating department, and dependence on local IT support.

The obvious alternative to locally distributed EDC systems – use of a central database server accessed via the Internet with browser software, employing hier-

<table>
<thead>
<tr>
<th>Variable</th>
<th>Significance/Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country, department</td>
<td>Identifier for benchmarking, comparisons</td>
</tr>
<tr>
<td>Date (year) of test</td>
<td>Seasonal effects, time trends</td>
</tr>
<tr>
<td>Patient identifier</td>
<td>Anonymous; link of multiple consultations by one patient</td>
</tr>
<tr>
<td>Consultation identifier</td>
<td>Anonymous; individual primary key for one consultation</td>
</tr>
<tr>
<td>Age, gender</td>
<td>Important gross exposure surrogates, used for direct standardisation (age usually dichotomised into &lt; 40 and &gt; 39)</td>
</tr>
<tr>
<td>Occupation (job title)</td>
<td>Exposure surrogate; ISCO-88 (starting 2014 in ESSCA: ISCO-08), partly aggregated to major/sub-major groups</td>
</tr>
<tr>
<td>Anatomical site(s)</td>
<td>Indicator of contact with allergen</td>
</tr>
<tr>
<td>Incriminated product</td>
<td>One or more product categories considered as cause of dermatitis; can be used as exposure surrogate</td>
</tr>
<tr>
<td>Allergens tested</td>
<td>Array of substances patch tested during consultation (multiple reading times of the test). The ‘baseline’ series tested in virtually all patients typically comprises 25 to 30 allergens, which are read 2 or 3, sometimes 4 times. Hence the matrix has 120 elements; considerably more if more allergens, such as up to 100, are being tested.</td>
</tr>
<tr>
<td>Allergens with (allergic) reactions</td>
<td>Subset of above, various reactions possible during the multiple reading times, usually aggregated to &quot;positive (allergic)&quot; vs. &quot;non-positive (not allergic)&quot; for common analyses. The complement has a negative result (&quot;not allergic&quot;).</td>
</tr>
<tr>
<td>Relevance of positive reactions</td>
<td>All allergic patch test reactions must at the final reading be assessed concerning their (unequivocal or probable) significance for current or past (occupational or non-occupational) allergic contact dermatitis</td>
</tr>
<tr>
<td>Final diagnosis/es</td>
<td>Allergic or irritant contact dermatitis, dermatitis of other or unknown aetiology, other types of allergic reactions, other dermatological diagnoses.</td>
</tr>
<tr>
<td>Occupational factors</td>
<td>Global statement whether current contact dermatitis is at least partially related to occupational exposures</td>
</tr>
</tbody>
</table>
architecturally controlled access to data – has hitherto not been implemented for contact allergy research in terms of a productive system, to the best of our knowledge. In other fields, with possibly less pronounced historical roots (which are definitely an obstacle for a uniform, consistent EDC solution), e.g. [6], this option has proven useful and viable.

The development and maintenance of locally installed software or a centrally hosted database necessitates specialised personnel and appropriate resources on a departmental and a central level, respectively. Hence, only (larger, adequately funded) networks will possibly be in a position to invest in such an undertaking. Single centres or smaller groups are advised to join existing networks after carefully checking evidence of previous productivity (list of publications using the system) and indicators of future viability.

3.2 ESSCA Data Centre

ESSCA became fully operational in 2002, with start-up funding by the EU not only for i) the data centre, but also ii) the development of the WinAlldat/ESSCA software [19] intended for consortium members not yet using a patch test software, or wanting to upgrade, and iii) local export interfaces for those network participants who continued to use their existing software. Initially, 17 departments from 9 countries were included, whereas the latest reporting period comprises 53 departments from 12 European countries [16, 19]. Since 2004, several new contributors to the ESSCA network have joined, necessitating (in lieu of export interfaces) the development of import scripts and mapping procedures upon receiving local data in their original shape in the data centre. The standard operating procedures of the ESSCA data centre are described in detail at http://www.essca-dc.org/doc/essca_central_SOPs_online.pdf (last accessed 2016-01-27). In brief, the following successive tasks are performed (at present, by one staff part-time):

1. Structured archiving of all incoming data in its original form, import as is into R using functions as appropriate.
2. Transformation into the central format of data storage, and mapping to centrally used catalogues, where necessary, using four data frames (relations) in a non-normalised relational database (‘test series’ data covering all information on patch test series used in the department, ‘case’ data pertinent to one consultation of a patient, such as age, sex, and clinical variables, ‘test’ data in terms of all patch test results, including negative, and ‘relevance’ data indicating the significance of a diagnosed contact allergy for the patient). All information is linked with appropriate keys.
3. Unless disjoint annual portions of finalised local data are provided, incoming data is incrementally added to existing data of that department, with strict precedence of newer over older data in case earlier data had been updated.
4. Data from one department collected in a particular year is used to prepare an “Internal Report”. The intention is to present to the local researcher a standard descriptive analysis of all relevant data, including the number and proportion of missing information. Should possible errors be spotted, the import (with new amended data and/or corrections of the import script) is re-done addressing these errors, and a new version of the “Internal Report” is prepared, until data are deemed valid and the annual portion of data can be added to the overall pool of data. The use of quality controlled data is believed to enhance acceptability of results by the scientific community. Technically, the internal report is a LaTeX file produced by the R function Sweave(). This enables to easily recycle a standard template, with both in-line, tabular and graphical results presentation and conditional commenting, using changing data portions from different departments and different years, respectively.
5. As soon as a sufficient body of pooled data has accumulated, the issue of coordination of requests from among the network participants for analyses and publications arises, to avoid duplicate or even competitive work. This includes a ‘call for co-authors’ by those leading a research topic. Contributors of data not opting for such a contribution are listed in an acknowledgement, for which consensus needs to be sought for each manuscript prior to submission.

The main scientific motivation for analysing patch test data is monitoring the frequency of contact allergy amongst patients – over time, across geographical regions, or in certain subgroups or concerning certain allergens. Given sufficient size a national network may be regarded as a clinical sample of the national level; some examples of surveillance are reviewed in [8] and [9]. General “good epidemiological practice” guidelines as well as guidelines more specific for the descriptive analysis and presentation of contact allergy research results must be considered [10 – 12].

4. Results

In this section, outcomes of the ESSCA network, both on an internal level, and in terms of informing the scientific community, are briefly presented. Another outcome, on a technical level, is the multilingual software WinAlldat/ESSCA [19] which is available for free, after registration, from the project’s website (www.essca-dc.org). Translations into languages other than English, Polish, Spanish and Italian can be made using a toolkit, provided upon request to the corresponding author. However, any (other) software solution can be used to contribute to the ESSCA network, provided it delivers structured data of adequate scope. The current spectrum of EDC systems used for contribution to ESSCA is shown in Figure 1. Presently, only the Slovenian network and the Information Network of Departments of Dermatology (IVDK, www.ivdk.org; comprising Austrian, German and Swiss departments) contribute multi-department data to the data centre, thereby rendering it to a meta-network; otherwise single departments submit their data individually.

The feedback to local departments in the standardised format of “internal reports” delivers a first result, as the patch test software used locally often provides only more limited capabilities of analysis – if these are used at all. Primarily intended as a quality control measure, to check for issues of data quality (missing data, plausi-
bility), or errors in the import or mapping process, the final “internal report” reflecting adequate data is appreciated as valuable information by the local colleagues. Moreover, a comparison with the general ESSCA results which are also provided (in the context of presentations at meetings or as publications, see below) allows some benchmarking of departmental results. Both the IVDK and the British Society of Cutaneous Allergy (BSCA, www.cutaneousallergy.co.uk) have formally integrated such auditing into their annual meetings, striving for standardisation and high quality data.

However, the main purpose of ESSCA is contact allergy surveillance and presentation of results at scientific congresses and as scientific publications. So far, data from 114,330 consultations have been collected between 2002 and 2014 from altogether 59 departments in 12 countries. Recent examples of the latter include a comparison of the time trend of contact allergy to nickel in four countries [20], an overview of the characteristics and selection of patients patch tested during the current reporting period 2009–2012 [16], patch test results in children and adolescents across Europe [21], or identification of high-risk occupations and important occupational allergens, respectively [22].

5. Discussion

We would not claim that the way the ESSCA network in general, and the data centre in particular, is organised should serve as a model for many or even all international epidemiological disease registries. In fact, other networks such as the “German Obstetric Surveillance System” [6] which do not need to incorporate diverse, historically grown EDC solutions, may find quite different solutions appropriate for their objectives. However, ESSCA has, from the start, not tried to impose such a “one solution fits all” EDC approach, but to offer maximum accessibility by providing locally installable software, and alternatively, also incorporating exports from existing departmental EDC solutions. Custom formatted exports, i.e. already in the necessary ‘common denominator format’, had been provided in the early period of the project (2002–2004) enabled by EU funding. Lately, however, exports are delivered to the data centre ‘as is’ and need to be appropriately transformed with custom import scripts. However, with adequate funding and input of IT knowledge (and support on the local level) much more consistent solutions along the lines of concepts outlined in the introduction may be developed. Still, we believe some technical aspects which have proven useful may be interesting for researchers organising a similar “investigator-initiated” network, namely

- a low threshold for joining the network by acceptance of all sufficiently structured data;
- responsibility for anonymisation directly at the department level and no transmission of personal identifiers to the data centre (the combination of age, sex and [aggregated] occupation is deemed insufficient to yield unique profiles within the catchment area of a department in case any central data is stolen and misused);
- structured feedback to participating departments in the format of “internal reports” which offers an incentive for participation in the network, in addition to the possibility to act as (co-) author of scientific publications;
- use of the Sweave() function provided by R which offers an elegant way of rationalising such reports, probably superior to components of other software packages, such as the “Output Delivery System” by SAS® in terms of adaptability;
- a clear organisation of processes to collect, pool, analyse and publish data which avoids uncertainties from the side of the network participants concerning their role.

Other aspects must be viewed as weaknesses or shortcomings, such as

- Some systems contributing to the ESSCA data pool do not allow the identification of re-consultations, i.e., a link between different consultations of one patient. According to data of the IVDK contact allergy network (where such re-identification is possible), the proportion of re-consultations is < 1% e.g. within a 9-year period [23] – and thus of impact only in long-term analyses, where the number of persons tested will be slightly inflated.
Since the ESSCA data collection platform was established, more comprehensive approaches towards decentralized data collection have been developed and published. Record linkage tools like the TMF PID Generator [24], the Mainzelliste [25], or the Mosaïc E-PIX service [26] provide methods for the centralized management of patient pseudonyms and allow to merge records when patients have been examined in multiple participating centres. The deployment of such a centralized ID management service, however, raises the barrier to entry in a distributed scenario with different data entry solutions in multiple countries that need to be integrated.

Metadata Repositories (MDRs) compliant to the ISO 11179-1 standard (http://metadata-standards.org/11179/, last accessed 2015-07-06) are being proposed to provide a shared structured representation of relevant research data items that can support the harmonisation of data acquisition across participating centres [27]. With open source implementations suitable for academic use recently becoming available (e.g. https://mdr.imise.uni-leipzig.de/) integration of an MDR should be considered for epidemiologic registries [28].

6. Conclusion

The ongoing work of national networks as well as international networks such as ESSCA has unequivocally proven the value of networking in this field, i) by establishing quality assurance measures (in single centre data, no benchmarking is possible) and ii) by providing important scientific evidence, e.g., on the recent, dramatic increase of contact allergy prevalence to the preservative methylisothiazolinone (e.g. [29–31]). Future solutions of (meta)network data collection in terms of disease registries should preferentially take up the more advanced, current solutions outlined in the discussion. The aim is a good (economic, scientifically valid, user-acceptable) balance between the effort for data entry locally and the scope and flexibility to cover routine surveillance and ad hoc research topics.

References