Standard operating procedures of the ESSCA data centre

As soon as several centres co-operate as a multicentre network, as in ESSCA, but also several national networks like the British BSCA (http://www.cutaneousallergy.org/, last accessed 2015-06-09) or the German/Austrian/Swiss IVDK (www.ivdk.org, last accessed 2015-06-09), the issue of handling data coming in from the different departments arises. The tasks of a data centre (for contact allergy surveillance) generally include the systematic steps outlined in the next sections. Names of R functions or packages, and of variables, are set in `Courier` font.

Contents

Check of incoming data

Data arriving in the different formats, both in terms of file type and structure, are stored in their original format in a folder denoting department and time of export. Each of these data files requires a custom import procedure, if possible (in case of text files) after conversion to UTF-8. All procedures are implemented in R (at the time of writing: version 3.1.3; www.r-project.org). To provide a full account of the data import and mapping and aggregating possibly involved, the respective R script is combined with comments and output of simple descriptive analyses in terms

of a protocol. The protocol is produced as LaTeX file by the R function `Sweave()`, combining a fixed layout and explanatory tables, program script chunks, short results tables and added comments, and compiled as pdf file. This is made available to the department delivering the data for plausibility checks.

- **Spreadsheets**, with personal identifiers of patients removed, offer first opportunities for checks (range of values, missing values). Subsequently, this data is converted into csv format and then imported into R using `read.csv()`, followed by further mapping as required to achieve the common denominator central format (Tables 1 - 4).
- Historically **DBF files** from a FoxPro database have been received from the department in Lahti, Finland. Import of the original relations was done using the `read.dbf()` function of the R package `foreign`.
- **MS-Access databases** exported from a local departmental database, with identifiers removed, as (multi-) annual subsets of that database. Using the `RODBC` package and an ODBC driver (user DNS specific for a certain portion of data) on a Windows ™ system, as the ODBC concept is not available under Linux, database relations are easily imported with the `sqlFetch()` function and stored as R data frames, representing identical copies of the original data. These are then merged and processed to eventually achieve the structure of the four meta-files.
- A set of exported data (“unloads”) from a relational database (such as “WinAlldat/ESSCA”) is received in csv format, as exact copies of the database tables. All csv files are imported with `read.csv()` and then processed further, as above.

Data structure used for pooled data

Generally, any consistent representation of the data can be used, provided it reflects well the structure of the data – to this end, the one which evolved for use in the ESSCA data centre is just

one of several options which has proven useful, similar also to the central data management of the IVDK (www.ivdk.org). Some limited redundancy in the data representation is intended to ease and speed up later analysis. The scope mainly of consultation related data reflects the 'minimal dataset', which had been decided by the ESSCA consortium.

Case/consultation data

A first data table contains demographic and clinical data related to the consultation of one patient; consultation comprising initial assessment, patch test (possibly in several successive time periods) and final evaluation. Hence, “a consultation” may cover several weeks, or just a few days. Some patch test software systems (such as WinAlldat/ESSCA or the sister software WinAlldat/IVDK used by the contributors of the IVDK) locally allow re-identification of a patient, if he or she consults another time. This is represented by a 1:n relation between “patient” and “case” numerical anonymous identifiers in the central data pool. However, it is generally not possible to identify repeat consultations across participating departments.

Table 1: ESSCA “minimal dataset”: case (consultation) related variables. Key variables set in bold face

Variable name	**Variable Type etc**	**Comments / Explanations**
center_id	Char(5)	2 letters with country code, hyphen, 2 digit number unique in country
pat_id	Integer	Should be unique per person
case_id	Integer	Should be unique per person*consultation
age	Integer	Calculated from date of consultation – date of birth or directly
sex	Char(1)	“M” or “F”
dateconsultation	Date	Date of consultation
test_date	Date	Date of patch test; derived from this: "tyear", see below
asthma	Char(1)	Y = yes, N = no, U = unknown
rhinoconjunctivitis	Char(1)	Y = yes, N = no, U = unknown
eczema	Char(1)	Y = yes, N = no, U = unknown
occup1	Integer	ISCO-88 (08, starting 2014) number, extended /

Variable name	Variable Type etc	Comments / Explanations
		condensed by ESSCA.
occup1_duration	Integer	Duration of occupation, in months
occup2	Integer	ISCO-88, see above
occup2_duration	Integer	see above
prev_occup1	Integer	ISCO-88, see above
prev_occup1_duration	Integer	see above
prev_occup2	Integer	ISCO-88, see above
prev_occup2_duration	Integer	see above
symptoms	Integer	Duration of dermatitis in months
pattern	Char(12)	A small catalogue of clinical patterns
primary_site	Integer	A numerically coded list of all anatomical sites, with 2-level hierarchy, according to catalogue table
secondary_spread	Char(1)	Y = yes, N = no, U = unknown
contactid_1	Integer	One of a list of 37 product types considered triggers of contact dermatitis, according to catalogue table
contactid_2	Integer	see above
contactid_3	Integer	see above
leisure_1	Integer	One of a list of categorised leisure activities considered causative for contact dermatitis, according to catalogue table
leisure_2	Integer	see above
leisure_3	Integer	see above
diag1	Integer	A numerically coded list of 51 diagnoses, relevant diagnoses further subdivided, , according to catalogue table
site1_diag1	Integer	Anatomical site for diag1, catalogue as for primary_site
site2_diag1	Integer	see above
site3_diag1	Integer	see above
diag2	Integer	See above ("diag1")
site1_diag2	Integer	Anatomical site for diag2, catalogue as for primary_site
site2_diag2	Integer	see above
site3_diag2	Integer	see above
diag_other_txt	Char(40)	free text for documenting “other” diagnosis
occ_factors	Char(1)	Occupational factors important for contact dermatitis: Y = yes, P = partly, N = no, U = unknown

Variable name	Variable Type etc	Comments / Explanations
endo_factors	Char(1)	Endogenous factors important for contact dermatitis: Y = yes, P = partly, N = no, U = unknown
exo_factors	Char(1)	Exogenous factors important for contact dermatitis: Y = yes, L = leisure activity, E = environmental, N = no, U = unknown
tyear	Integer	Year of patch test (from test_date)

Test series data

The patch test is traditionally not performed by applying a varying, completely individualised set of potential allergens to each single patients, but organised by using so-called “test series”. These are collections of substances which have proven important for virtually every patient (these are assembled in a 'baseline series' normally applied to each patient patch tested) and those substances which are important only in certain exposures, such as 'hairdressing chemicals', 'rubber additives', etc. All test series are periodically adapted to changing requirements (allergens which are no longer important are phased out, while “new” allergens are introduced, or test concentrations or vehicles are adapted to new evidence). Hence, a certain “version” of a test series valid at that time is applied to a patient patch tested at a given point in time. Moreover, different departments may introduce (slightly) different test series adapted to their purposes and research interests. To achieve historical correctness, all versions of test series ever used in the patients are included in a second data table (Table 2).

Table 2: ESSCA “minimal dataset”: test series related variables. Key variables set in bold face

Variable name	Variable Type etc.	Comments / Explanations
center_id	Char(5)	See above
series_id	Integer	Unique identifier of one version of a test series of allergens
series_nr	Integer	Number of the test series (arbitrary)
ser_name	Char(30)	Free text name for (version of) test series

Variable name	Variable Type etc.	Comments / Explanations
method	Char(1)	A = patch test (other values irrelevant)
contactid	Integer	See above ("contactid_1"). Here, an attribute of a test series
standard_	Integer	Flag: 1 for testing in consecutive patients (“baseline series”)
valid_from	Date	Begin of use of the version of the test series
valid_to	Date	End of use of the version of the test series
position	Integer	Rank order of the allergen in the series (1 to n)
substanceid	Integer	Unique central number for a certain substance, according to catalogue table. This often needs mapping / amendment steps in data pooling.
substancename	Char(40)	Substance main name
concentration	Real	The patch test concentration
vehicle	Char(3)	Shorthand for test vehicle, such as “PET” for petrolatum, “AQU” for water, according to catalogue table
maker	Char(2)	Shorthand for supplier of patch test material, according to catalogue table

Test result data

The actual results with a given test series, or several series, obtained at the consultation of a patient, are stored in another data table (Table 3). Its key variables link to consultation data (via center_id, pat_id, case_id, see Table 1) and to test series data (via center_id, series_id, see Table 2). If the entire test series did not elicit any reaction, having been completely tested (sometimes certain substances are omitted for different reasons), the value of the variable position remains blank and a zero in those days when readings had been performed is inserted (typically d2=0, d4=0). As soon as either any reaction to a substance in a test series is observed at any of the multiple reading times, or a substance had been omitted, position becomes a part of the key, referring to the respective row of test series data identified by center_id, series_id and position. Again, the reading times used are represented by the values inserted, e.g., d2=0, d4=45 (negative on day 2, strong positive reaction, according to a catalogue table, on day 4). This implies that all other allergens of the same test series were negative

(0) at D2 and D4. Thereby, a compact storage of all necessary information to reconstruct reactivity to any allergen, at any reading time in the time frame d1 to d7, during later analyses can be achieved.

Table 3: ESSCA “minimal dataset”: patch test related variables. Key variables set in bold face

Variable name	**Variable Type etc**	**Comments / Explanations**
`center_id`	Char(5)	See above (Table 1)
`pat_id`	integer	See above (Table 1)
`case_id`	integer	See above (Table 1)
`series_id`	integer	See above (Table 2)
`position`	Integer	See above (Table 2)
`d1`	Integer	Numeric coding for patch test reaction, day 1, according to catalogue table. E.g., 0 is neg, 22 is doubtful, 37 is +, 45 is ++, 50 is +++, and 1 to 3 are “not tested” (important, the latter needs to be subtracted from the denominator of tested patients)
`d2`	Integer	Numeric coding for patch test reaction, day 2, see above
`d3`	Integer	Numeric coding for patch test reaction, day 3, see above
`d4`	Integer	Numeric coding for patch test reaction, day 4, see above
`d5`	Integer	Numeric coding for patch test reaction, day 5, see above
`d6`	Integer	Numeric coding for patch test reaction, day 6, see above
`d7`	Integer	Numeric coding for patch test reaction, day 7, see above
`dxfinal`	integer	Numeric coding for maximum patch test reaction between day 3 and day 5 (inclusive), see above
`dxsum`	Integer	Numeric coding for maximum patch test reaction, day 1 to day 7, see above

Certain conventions apply for data representation in above structure:

- If certain readings on D1 to D7 are not performed, the respective value is NA (“missing”)
- If none of the allergens of a test series yielded any reaction, and all allergens had actually been applied, there is one data row with `center_id`, `pat_id`, `case_id`, `series_id` as key variables, `position` NA, and the value(s) 0 in "`d1`" through "`d7`", as appropriate (as read)
- If at least one allergen of a test series yielded a reaction (or was not tested), this allergen is represented in one data row with `center_id`, `pat_id`, `case_id`, `series_id`, `position` as key variables, and the appropriate values in `d1` through `d7` (as actually read). In this case, the other allergens of the test series (unless they also yielded a reaction or were actually not tested and are thus represented by another data row) are implicitly regarded as negative, at the reading times represented in the present data row.

Data on “clinical relevance”

Upon final evaluation, at the conclusion of the patch test, several diagnostic variables are generated, including information represented in the consultation data (Table 1) starting with `diag1`.

Moreover, positive (allergic) reactions observed to the set of allergens applied to the patient are normally assessed concerning their “clinical relevance” (1). Clinical relevance describes whether the contact allergy to a certain substance “explains” either a current, or a previous episode of contact dermatitis. Often, relevance is separately considered for occupational and non-occupational contacts. Sometimes, a statement on the likelihood of the statement is added, yielding categorisation such as “certain current non-occupational relevance” (e.g., of contact allergy due to chromium, causing dermatitis of the feet induced by chromium-tanned leather shoes) or “probable past occupational relevance” (e.g., of contact allergy to a biocide which is often found in cutting fluids, probably explaining work-related contact dermatitis of the hands in a patient previously working as machinist).

Such data is represented in Table 4.

Table 4: ESSCA “minimal dataset”: variables related to clinical relevance. Key variables set in bold face

Variable name	**Variable Type etc**	**Comments / Explanations**
`center_id`	char(5)	See above (Table 1 and 3)
`case_id`	Integer	See above (Table 1 and 3)
`substanceid`	Integer	See above (Table 2)
`method`	Char(1)	See above (Table 2)
`relevanceid`	Integer	Numerical representation of a 3-level hierarchical catalogue concerning clinical relevance
`relevance_txt`	Char(30)	Textual representation of above (redundant)
`contactid`	Integer	See above (Table 1: "`contactid_1`") Used here to qualify which kind of product has caused the reaction to the allergen in the patient

Merge with existing data from a department

While all data portions are preserved at the data centre in the very state they had been delivered for archiving purposes, new data in terms of updates of already existing data arriving with a new delivery are, as a convention, regarded as overriding existing data. Accordingly, the cumulative data contribution of one department is (i) supplemented with data from consultations of new patients and (ii) updated by edits of existing data provided with the delivery. This update/addition is performed on the level of the central “common denominator” data structure, no matter the underlying original data structure. Corrections typically affect only cases having consulted in the period preceding the currently delivered period. If data are delivered as strictly annual, disjunct portions of finalised data, the issue of merging and updating evidently does not arise.

“Internal reports” for quality control

An “internal report” on the data provided for each one year, by one department, is produced immediately after importing data, and send to the delivering department along with the annotated import script mentioned above, evidently preceded by scrutiny, and revision cycles as indicated, at the data centre. 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 amount of missing data (if in a problematic level of e.g. > 10 % for core data such as age and gender and > 20% for other data, as defined by the peer group (2)) can sometimes not be amended in retrospect, but needs to be addressed prospectively. 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.

Beyond the primary intention, “internal reports” in their final, definite version can also address quality control as a key part of the process of clinical governance. Specifically, each scientific network analysing pooled data should be aware of the variation introduced by methodological or technical differences (patch test material, selection of test series to be applied and, possibly most crucial, visual scoring of the test reactions (3)). Primarily the comparison of local results (the % positive reaction to a set of allergens) with the group's average is used to identify possible methodological variation (4). Once identified, efforts should follow to eliminate the “technical” source(s) of divergence. However, differences in the spectrum of allergens between the centres can also be due to differences in patient characteristics, e.g. described by the so-called MOAHLFA-index (5). This incorporates the most important demographic and clinical characteristics known to be associated with many contact allergies, helping to put 'outlying' patch test results into this perspective (6). While it is unlikely that any single measure of quality will indicate a cause for concern with regard to local methodology and interpretation of results, striking differences, particularly when repeated in data pooled over 2-3 years, should lead to careful reflection as to the reasons (4; 7).

Coordination of requests for analyses and publications

With accumulating pooled data, options for scientifically relevant analyses increase beyond routine reporting of basic results. Requests for data analyses from network participants must be co-ordinated 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 consent needs to be sought for each manuscript prior to submission. The response to possible requests for data analyses and interpretation from outside the network (e.g., industry associations, regulatory institutions) is managed by a board of directors elected from the general ESSCA assembly in Berlin, June 2004.

As a special case of request for analysis, departmental data can be delivered back to the department in a format which makes the data amenable to autonomous statistical analysis by the department's researchers. For this purpose, data is typically transformed and aggregated to the “flat” spreadsheet structure which is used by some participants to deliver data, because even without data management skills this can be descriptively analysed with spreadsheet software already, or imported into various statistical software used at the department and analysed further.

Analysis of quality-controlled pooled data

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) (11) (12).

References

(1) Johansen J D, Aalto-Korte K, Agner T, Andersen K E, Bircher A, Bruze M, Cannavó A, Giménez-Arnau A, Gonçalo M, Goossens A, John S M, Lidén C, Lindberg M, Mahler V, Matura M, Rustemeyer T, Serup J, Spiewak R, Thyssen J P, Vigan M, White I R, Wilkinson M, Uter W. European Society of Contact Dermatitis guideline for diagnostic patch testing - Recommendations on best practice. *Contact Dermatitis* 2015: **73**: 195-221.

(2) Uter W, Mackiewicz M, Schnuch A, Geier J. Interne Qualitätssicherung von Epikutantest-Daten des multizentrischen Projektes Informationsverbund Dermatologischer Kliniken (IVDK). *Derm Beruf Umwelt* 2005: **53**: 107-114.

(3) Svedman C, Isaksson M, Björk J, Mowitz M, Bruze M. 'Calibration' of our patch test reading technique is necessary. *Contact Dermatitis* 2012: **66**: 180-187.

(4) Britton J E R, Wilkinson S M, English J S C, Gawkrodger D J, Ormerod A D, Sansom J E, Shaw S, Statham B. The British standard series of contact dermatitis allergens: validation in clinical practice and value for clinical governance. *Br J Dermatol* 2003: **148**: 259-264.

(5) Schnuch A, Geier J, Uter W, Frosch P J, Lehmacher W, Aberer W, Agathos M, Arnold R, Fuchs T, Laubstein B, Lischka G, Pietrzyk P M, Rakoski J, Richter G, Ruëff F. National rates and regional differences in sensitization to allergens of the standard series. Population-adjusted frequencies of sensitization (PAFS) in 40,000 patients from a multicenter study (IVDK). *Contact Dermatitis* 1997: **37**: 200-209.

(6) Uter W, Gefeller O, Geier J, Schnuch A. Changes of the patch test population (MOAHLFA index) in long-term participants of the Information Network of Departments of Dermatology, 1999-2006. *Contact Dermatitis* 2008: **59**: 56-57.

(7) Bourke J, Coulson I, English J. Guidelines for the management of contact dermatitis: an update. *Br J Dermatol* 2009: **160**: 946-954.

(8) Schnuch A, Geier J, Lessmann H, Arnold R, Uter W. Surveillance of contact allergies: methods and results of the Information Network of Departments of Dermatology (IVDK). *Allergy* 2012: **67**: 847-857.

(9) Uter W, Schnuch A, Giménez-Arnau A, Orton D, Statham B. Databases and Networks. The Benefit of Research and Quality Assurance in Patch Testing.. In: Contact Dermatitis, 5. edition, Johansen, J.; Frosch, P. & Lepoittevin, J.-P. (Eds): , Springer, 2011 : 1053-1064.

(10) Uter W, Schnuch A, Gefeller O. Guidelines for the descriptive presentation and statistical analysis of contact allergy data. *Contact Dermatitis* 2004: **51**: 47-56.

(11) Gefeller O, Pfahlberg A B, Uter W. What can be learnt from nothing? A statistical perspective. *Contact Dermatitis* 2013: **69**: 350-354.

(12) Uter W, Rustemeyer T, Wilkinson M, Johansen J D. Quality in epidemiological surveillance of contact allergy *Contact Dermatitis* 2016: **(accepted)**.