UniversitätsKlinikum Heidelberg

Privacy and Data Coarsening: A Clinical Trials Perspective

Johannes Hüsing

Koordinierungszentrum für Klinische Studien, Universität Heidelberg

TMF-Workshop, 2015-03-19

Outline

Cross-Study Databases, Sharing of Data in Theory

- Cross-Study Information
- EMA Policy/0070: Transparency in Clinical Research

Practical Implementation (Industry)

Conclusions

Identification subjects within trials

Clinical trials address single research questions, sample size and CRF tailored to specific needs.
Study data are pseudonymized.
Code is retained with investigator:

- Retrieve patient file for repeated measurements within trial
- Within audit, demonstrate that subjects actually exist

Code list must be retained for “an agreed on time” (ICH E6).

Linking subjects across trials

Especially in chronical diseases, a subject may incidentally have participated in multiple trials.
In individual patient data meta-analyses, these cases are largely ignored.
Code lists may not be generally used for linking.
However, FDA requires that

> *Subjects that participate in more than one study should maintain the same USUBJID across all studies [within one submission]. It is important to follow this convention to enable pooling of a single subject's data across studies (e.g., a randomized control trial and an extension study).*

The example does not explicitly address "incidental" matches, but the problem is taken seriously (PhUSE working group on "USUBJID best practice") within the industry.

Publication and access to clinical-trial data

- EMA policy paper, draft status as of 2013-06-24
- Open clinical trial data

Policy Goals

- *benefit public health in future*
- *level playing field that allows all drug developers to learn from past successes and failures*
- *enable the wider scientific community to make use of detailed and high-quality CT data*
- *take regulatory decision-making one step closer to EU citizens and patients*

Protection of Personal Data, Usage of Data

- *is a fundamental right of EU citizens*
- *policy ... must be fully compliant with applicable regulations in the EU, in particular Regulation (EC) No 45/2001 and Directive 95/46/EC*
- *concerned that emerging technologies for data mining and data base linkage will increase the potential for unlawful retroactive patient identification*
- *preventing rare but potentially damaging instances of patient identification*
- *Patients participate ... in the hope that their data will support the development and assessment of ... the treatment of their disease ... any other use of patient data oversteps the boundaries of patients' informed consent*

Protection of commercially confidential information

- *CT data cannot be considered commercially confidential information*
- *the interests of public health outweigh considerations of commercially confidential information*

Protection of commercially confidential information

- *CT data cannot be considered commercially confidential information*
- *the interests of public health outweigh considerations of commercially confidential information*

... it would be interesting ... to know whether the well written draft of EMA policy/0070 is aiming at maximal transparency simply because it is expected to be cut down later anyway.

Koenig et al, *Biometrical Journal* **57** (2015) 1, 8–26 DOI: 10.1002/bimj.201300283

Measurements to ensure protection of personal data

Appropriate de-identification Recommended minimum standard: Hrynaszkiewicz, BMJ 2010; ...

should be such that adherence will preclude subject de-identification, even when applying linkages with other data carriers (e.g. social media)

Controlled access

- requester identifies herself
- requester (natural or legal person) established in the EU
- analysis solely for public health, including exploratory analysis for hypothesis generation, aims are disclosed
- no attempt to reidentify patients
- no usage of data outside patients' informed consent

Measurements to ensure protection of personal data

Appropriate de-identification Recommended minimum standard: Hrynaszkiewicz, BMJ 2010; ...

should be such that adherence will preclude subject de-identification, even when applying linkages with other data carriers (e.g. social media)

Controlled access

- requester identifies herself
- requester (natural or legal person) established in the EU
- analysis solely for public health, including exploratory analysis for hypothesis generation, aims are disclosed
- no attempt to reidentify patients
- no usage of data outside patients' informed consent

Scope of informed consent?

> *"... dass eine allgemein formulierte Zweckbestimmung und eine Einwilligung in eine offenere spätere Nutzung ... immer nur dann eine ausreichende Rechtsgrundlage für die Forschung bieten können, wenn die Einholung späterer Einwilligungen ... nicht machbar ist."*
>
> *Pommerening et al. Leitfaden zum Datenschutz in medizinischen Forschungsprojekten. Schriftenreihe der TMF, Berlin 2014*

Hrynaszkiewicz et al, BMJ 2010: Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers

- Benefits of sharing clinical trial data along with publication are highlighted
- List of 28 possibly identifying variables, following publications of numerous bodies (National Information Governance Board, Medical Research Council, Department of Health (NHS Code of Practice), UK Data Archive, International Committee of Medical Journal Editors, Washington State Department of Health, Partners Human Research Committee (HIPAA frequently asked questions), National Heart Lung and Blood Institute)

List of 28 variables, Part I: direct

- Name, initials, address (postal code), telephone/fax numbers, electronic mail addresses, names of relatives, unique identifying numbers
- Vehicle identifiers, medical device identifiers
- Biometric data, facial photograph or comparable image, audiotapes
- Date of birth, other dates including day of treatment

List of 28 variables, Part II: indirect

- Rare disease or treatment
- Sensitive data, such as illicit drug use or “risky behaviour”
- Place of birth, socioeconomic data, such as occupation or place of work, income, or education, household and family composition, ethnicity
- Anthropometry measures, multiple pregnancies
- Small denominators—population size of < 100, very small numerators—event counts of < 3
- Year of birth or age
- Verbatim responses or transcripts

Observations on item list

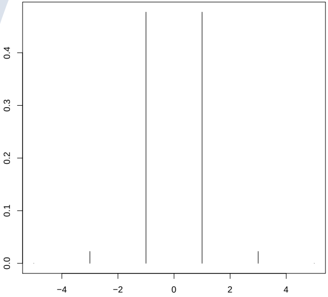
- Most of the critical variables are individual traits, which need to persist for a longer time to deanonymize a subject
- Clinical trials care about the effect of an intervention on an outcome.
- Interventions are often the result of “coin tosses” and do not entail individual information at all, provided treatment does not become a “personal trait”
- Outcomes are often ephemeral (eg laboratory values). Sometimes they might be persistent, like a heart condition or permanent disability

⇒ Generally, most data critical to the analysis are less critical for identification.

Effect of coarsening on power

- Coarsening the response (relevant for clinical trials)
- Coarsening an explanatory variable (less relevant)
- Coarsening both (worst case scenario)

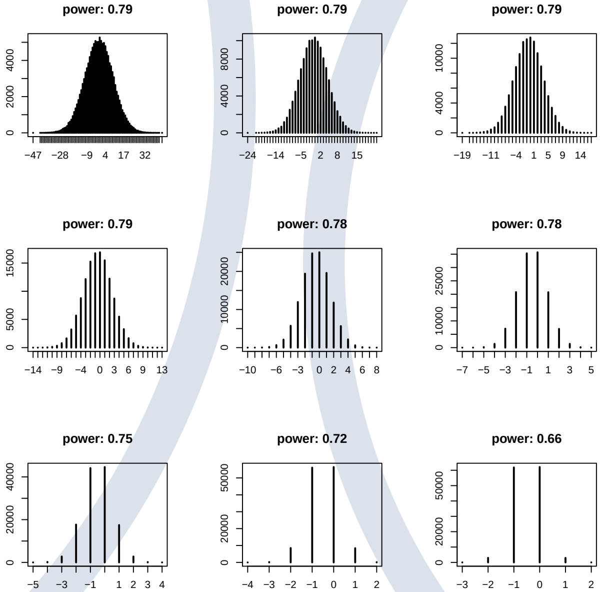
Rounding the response: two-sample t-test



- Tricker (1990) simulated sample sizes of 5, 10, and 25 per group
- Acceptable rounding bins were as wide as 2 standard deviations

Rounding response and explanatory variables

Rounding response and explanatory variables



Treating date values

In circumstances where it is essential for the scientific validity of the study to include dates, such as dates of treatment (a direct identifier), data must be presented in such a way that is unlikely to affect statistical analyses but preserves anonymity. For example, one could add or subtract a small, randomly chosen number of days to all dates, so that the true dates are not published.

Treating date values

In circumstances where it is essential for the scientific validity of the study to include dates, such as dates of treatment (a direct identifier), data must be presented in such a way that is unlikely to affect statistical analyses but preserves anonymity. For example, one could add or subtract a small, randomly chosen number of days to all dates, so that the true dates are not published.

Similar to Caesar cipher.

Relative dates and times fair game

- Timing of medication, biological reactions and measurements is essential in clinical trials
- Deducting date of randomization or first treatment from all dates preserves information
- Doesn't open whole data set to Caesar cipher attack, only subject by subject
- Minor caveat: Relative date values in CDISC SDTM are stored in a format

  > *... not suited for use in subsequent numerical computations, such as calculating duration. The raw date values should be used rather than Study Day in those calculations.*
  >
  > SDTM Implementation Guide 3.1.4

- Another minor caveat: Sometimes thr trial outcome is affected by general trend. Not possible to investigate once dates are lost (one could retain randomization sequence as a compromise)

Outline

Cross-Study Databases, Sharing of Data in Theory

- Cross-Study Information
- EMA Policy/0070: Transparency in Clinical Research

Practical Implementation (Industry)

Conclusions

Data brokering site

ClinicalStudy DataRequest.com `https://www.clinicalstudydatarequest.com`

- Run by several pharma companies
- Handles requests as part of research projects
- Offers access to study data via remote desktop

Controlled access part

- Applicant provides research proposal
- Independent review panel considers proposal
- Data sharing agreement

Research proposal

This is in line with EMA Transparency policy:

> *... the Agency considers preparation and uploading of a detailed protocol/statistical analysis plan before data access of utmost importance, to ensure the credibility of subsequent results*

EMA Policy/0070

Appropriate use of study data

From the Clinical Study Data Request user guide:

> *Please note that to ensure the use of the data aligns with the informed consent provided by clinical study participants, research proposals must relate to the medicine or disease that was the subject of the original study.*

Appropriate use of study data

From the Clinical Study Data Request user guide:

> *Please note that to ensure the use of the data aligns with the informed consent provided by clinical study participants, research proposals must relate to the medicine or disease that was the subject of the original study.*

EMA Policy:

> *Patients participate ... in the hope that their data will support the development and assessment of ... the treatment of their disease ...*

Actual access to study data

ClinicalStudy DataRequest.com

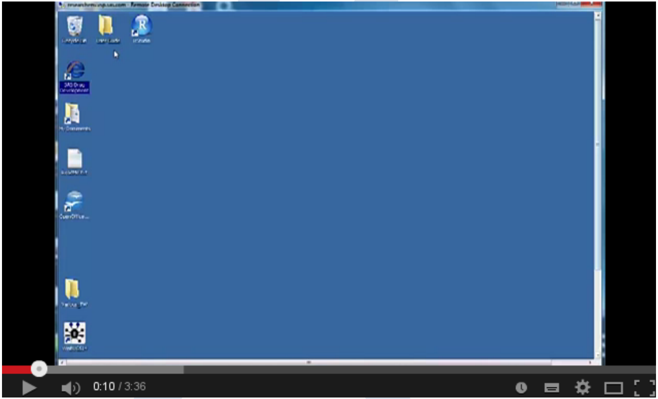
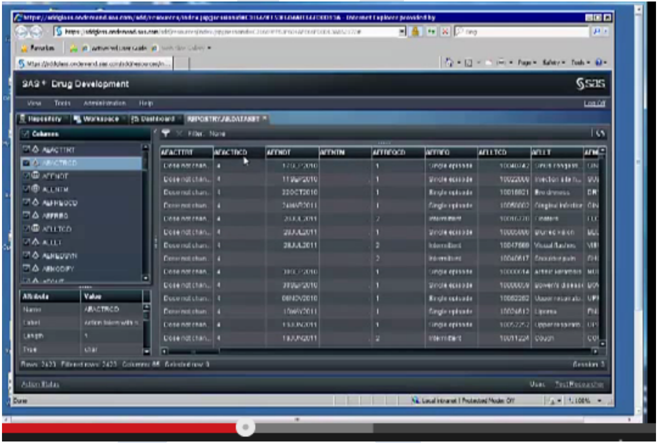


Table view on data

ClinicalStudy DataRequest.com



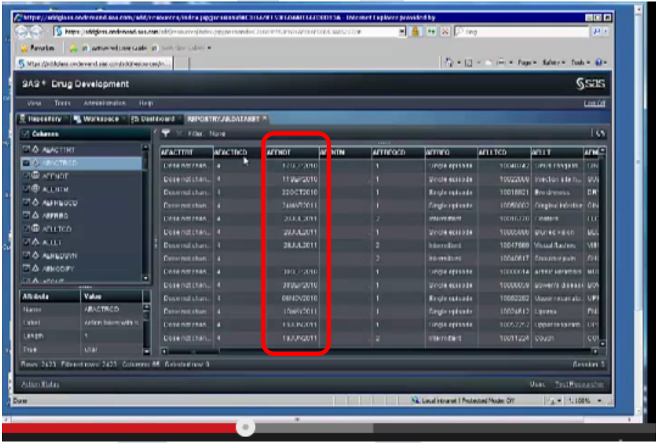
Attribute	Value
Name	AEACTRCD
Label	Action taken with s...
Length	1
Type	char

AEACTTRT	AEACTRCD	AEENDT	AEENTM	AEFREQCD	AEFREQ	AELLTCD	AELLT
Dose not chan...	4	17OCT2010		1	Single episode	[illegible]	[illegible]
Dose not chan...	4	11MAR2010		1	Single episode	[illegible]	Injection site ...
Dose not chan...	1	22OCT2010		1	Single episode	10016821	[illegible]
Dose not chan...	4	24MAR2011		1	Single episode	[illegible]	Fungal infection
Dose not chan...	4	20JUL2011		2	Intermittent	[illegible]	[illegible]
Dose not chan...	4	20JUL2011		1	Single episode	[illegible]	Blurred vision
Dose not chan...	1	28JUL2011		2	Intermittent	10047680	Visual flashes
Dose not chan...	4			2	Intermittent	10040617	Shoulder pain
Dose not chan...	4	30OCT2010		1	Single episode	[illegible]	[illegible]
Dose not chan...	4	29MAR2010		1	Single episode	[illegible]	[illegible]
Dose not chan...	1	08NOV2010		1	Single episode	10062282	Upper respirato...
Dose not chan...	4	10MAY2011		1	Single episode	[illegible]	Lipoma
Dose not chan...	4	15JUN2011		1	Single episode	[illegible]	Upper respirato...
Dose not chan...	4	18JUN2011		2	Intermittent	10011224	Cough

Rows: 2423 Filtered rows: 2423 Columns: 88 Selected row: 3

Appropriate de-identification?

ClinicalStudy DataRequest.com



AEACTTRT	AEACTRCD	AEENDT	AE?NTM	AEFREQCD	AEFREQ	AELLTCD	AELLT
Dose not chan...	4	17OCT2010		1	Single episode	10040742	Sinus congesti...
Dose not chan...	4	11MAR2010		1	Single episode	10022000	Injection site h...
Dose not chan...	1	22OCT2010		1	Single episode	10018821	Eye dryness
Dose not chan...	4	24MAR2011		1	Single episode	10050002	Gingival infection
Dose not chan...	4	20JUL2011		2	Intermittent	[illegible]	[illegible]
Dose not chan...	4	28JUL2011		1	Single episode	10005886	Blurred vision
Dose not chan...	1	28JUL2011		2	Intermittent	10047688	Visual flashes
Dose not chan...	4			2	Intermittent	10040617	Shoulder pain
Dose not chan...	4	30OCT2010		1	Single episode	[illegible]	[illegible]
Dose not chan...	4	31MAR2010		1	Single episode	[illegible]	Bowen's disease
Dose not chan...	1	08NOV2010		1	Single episode	10062282	Upper respirato...
Dose not chan...	4	10MAY2011		1	Single episode	[illegible]	Lipoma
Dose not chan...	4	15JUN2011		1	Single episode	10052252	Upper respirat...
Dose not chan...	4	18JUN2011		2	Intermittent	10011224	Cough

Columns: AEACTTRT, AEACTRCD, AEENDT, AEENTM, AEFREQCD, AEFREQ, AELLTCD, AELLT, AEMEDDRV, AEMODIFY, AEOUT

Attribute	Value
Name	AEACTRCD
Label	Action taken with s...
Length	1
Type	char

Rows: 2423 Filtered rows: 2423 Columns: 48 Selected row: 3 Session: 3

KKS Heidelberg Koordinierungszentrum für Klinische Studien

Outline

Cross-Study Databases, Sharing of Data in Theory

- Cross-Study Information
- EMA Policy/0070: Transparency in Clinical Research

Practical Implementation (Industry)

Conclusions

The situation

Clinical trial data are unaffected (exceptions apply) by some privacy concerns:

- Subject identification lists (“code lists”) may be destroyed after “an agreed on time” (ICH E6)
- Subject linking across studies is generally neglected (at least in academia)
- Most interesting data are transient and don’t identify patients

However:

- Clinical trials data sets are small (k-anonymization within a single data set futile)
- They are built by small research groups (no data sharing infrastructure)
- Small differences between informed consent forms

What is needed

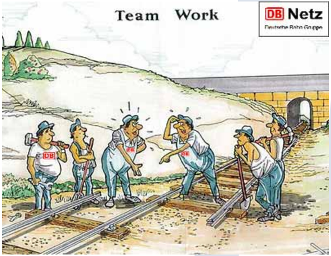
A clinical trial data repository that ...

- last longer than the trial
- account for individual informed consents (secondary research on disease? treatment? both?)
- coarsen data in a form that doesn't substantially affect power
- controls access based on pre-specified research concepts and analysis plans
- shares responsibility for data protection with applicant

As data collection in clinical trials is highly regulated, a "one size fits all" approach for a repository seems feasible.
In an online survey, 80 % of the Cochrane Collaboration's IPD Meta-analysis Methods Group favored a central data repository.

Tudur Smith et al, *Trials* **12** Suppl. 1 (2011),
doi:10.1186/1745-6215-12-S1-A56

The rest is merely implementation details



Age (a)			
Trial A		Trial B	
$18 \le x < 20$	3	$18 \le x < 25$	7
$20 \le x < 30$	9	$25 \le x < 35$	12
$30 \le x < 40$	13	$35 \le x < 45$	9
$40 \le x < 50$	11	$45 \le x < 55$	18
$50 \le x < 60$	26	$55 \le x < 65$	19
$60 \le x < 70$	22	$65 \le x < 75$	23