

Estimation of Patient Accrual Rates in Clinical Trials Based on Routine Data from Hospital Information Systems

M. Dugas¹; S. Amler¹; M. Lange²; J. Gerß¹; B. Breil¹; W. Köpcke¹

¹Department of Medical Informatics and Biomathematics, University of Münster, Münster, Germany;

²IT Centre, Universitätsklinikum Münster, Münster, Germany

Keywords

Patient accrual rate, hospital information system, clinical trial

Summary

Background: Delayed patient recruitment is a common problem in clinical trials. According to the literature, only about a third of medical research studies recruit their planned number of patients within the time originally specified.

Objectives: To provide a method to estimate patient accrual rates in clinical trials based on routine data from hospital information systems (HIS).

Methods: Based on inclusion and exclusion criteria for each trial, a specific HIS report is

generated to list potential trial subjects. Because not all information relevant for assessment of patient eligibility is available as coded HIS items, a sample of this patient list is reviewed manually by study physicians. Proportions of matching and non-matching patients are analyzed with a Chi-squared test. An estimation formula for patient accrual rate is derived from this data.

Results: The method is demonstrated with two datasets from cardiology and oncology. HIS reports should account for previous disease episodes and eliminate duplicate persons.

Conclusion: HIS data in combination with manual chart review can be applied to estimate patient recruitment for clinical trials.

Hospital information systems (HIS) contain data items, which are relevant for inclusion and exclusion of patients to clinical trials. For instance, diagnosis information is coded in HIS for billing purposes, but can also be analyzed to screen for potential trial subjects [4].

However, electronic patient records contain a lot of unstructured text information, therefore automated data analysis has limitations and expert review of records is needed to assess patient eligibility. In this context, we propose a method to estimate patient accrual rates based on HIS reports in combination with manual review of a sample of HIS records.

Methods

Because not all information relevant for assessment of patient eligibility is available as coded HIS data items, a two-stage process to estimate patient accrual rates is applied: First, a list of matching patients is generated with a specific HIS report for a given time span T (for instance, $T = [\text{January 1, 2007; December 31, 2007}]$). Second, a sample of these patient records is reviewed manually by an expert to assess eligibility and thereby estimate patient accrual rate.

HIS reports are database queries which can be generated using reporting tools of the HIS (HIS report generator) or by data queries from a data warehouse. These reports can access all structured data elements within the HIS. Typical examples of HIS data items are admission and discharge diagnoses (primary as well as secondary diagnoses, coded according to international classification of diseases), patient age, patient gender and routine lab values. Depending on inclusion and exclusion criteria of each trial, all suitable HIS items should be considered for this HIS report to provide high recall and precision.

Methods Inf Med 3/2009

Correspondence to:

Prof. Dr. Martin Dugas
Department of Medical Informatics and
Biomathematics
University of Münster
Domagkstraße 5
48149 Münster
Germany
E-mail: dugas@uni-muenster.de

Methods Inf Med 2009; 48: 263–266

doi: 10.3414/ME0582

received: June 4, 2008

accepted: November 26, 2008

prepublished: March 31, 2009

Introduction

Delays in patient recruitment are a common problem in clinical trials. Charlson [1] analyzed trials listed in the 1979 inventory of the National Institute of Health. He found that only 14 of 38 (37%) trials reached planned recruitment. Twenty-three years later a review of 114 trials between 1994 and 2003 held by the Medical Research Council and Health Technology Assessment Programmes found that less than one-third recruited their original target within the time originally specified [2]. There is a variety of reasons, such as fewer

patients eligible than expected, staff problems, limited funding, complexity of trial design, length of recruitment procedure and others. A recent Cochrane review [3] analyzed strategies to improve recruitment to research studies. Monetary incentives, an additional questionnaire at invitation and treatment information on the consent form demonstrated benefit; the authors concluded that these specific interventions from individual trials are not easily generalizable.

Therefore from a methodological point of view, methods are needed to estimate patient accrual rates in clinical trials more precisely.

HIS documentation is focused at a “case”, i.e. a certain episode of care in a hospital with related clinical and administrative data; trials are addressing individual patients. For this reason HIS reports for patient accrual should analyze all HIS cases of a patient to avoid duplicate persons and to account for pre-existing diseases. We propose a stepwise approach for those HIS reports: First, select all HIS cases matching inclusion and exclusion criteria; second, remove duplicate persons; third, identify all HIS cases for each matching patient and retrieve data on pre-existing diseases to check inclusion and exclusion criteria for each patient. For instance, many trials recruit patients with initial diagnosis, therefore it needs to be verified whether this diagnosis was established in the past.

Output of this report should be pseudonymized to protect patient data. The number of patients on this HIS report for time span T is denoted as n_T . Under the assumptions that average patient accrual rate does not change over time and the HIS report identifies exactly all eligible patients, estimated patient accrual rate would be $n_T/|T|$, where $|T|$ denotes the length of time span T .

However, typically only a subset of information required for inclusion and exclusion is available as coded HIS data items. Therefore only a subset of n_T matches all inclusion and exclusion criteria for a specific trial. A manual expert review of a sample with s_T patient records from the HIS report results in m_T matching patients.

Manual review of HIS patient records requires access to identifiable patient data, therefore it needs to be compliant with data protection laws. Physicians with direct involvement into patient care are allowed to access records of their patients. Therefore these physicians get a list of pseudonyms from the HIS report, which enables them to access those patient records. They report for each pseudonym, whether this patient is eligible for the trial without disclosure of the person's identity. In general, data access policies must be approved by the responsible data protection officer.

Before patient accrual rate is estimated, we propose to assess, whether the probability of HIS patients actually matching to the trial is constant over time. Therefore the number of matching and non-matching patients is figured out in a contingency table for a set of predefined sub-intervals t of time span T . Our null hypothesis states that the proportion of matching patients among all reviewed sample patients m_t/s_t is constant for all sub-intervals t and is tested by Pearson's Chi-squared test. If the null hypothesis is not rejected ($p > 0.05$), we conclude that the probability of HIS patients matching to the trial is constant in time and estimate patient accrual rate (PAR) in the total time span T as follows:

$$\text{PAR} = (m_T/s_T) * (n_T/|T|) \quad (1)$$

A confidence interval for the expected PAR can be calculated according to Clopper [6], as

implemented in R-function `binom.test` [5]. Specifically, we assume a fixed rate $n_T/|T|$. The supposed rate is multiplied by the calculated confidence interval of the probability of HIS patients actually matching to the trial.

Results

We use datasets from ongoing Münster atrial fibrillation trials [7] and leukemia trials [8,9] to demonstrate this method of patient accrual rate estimation. A HIS-based notification system generated HIS reports for study physicians, who manually reviewed patient records to assess trial eligibility [4].

Example 1: Atrial Fibrillation Trial

► Table 1 presents number of patients identified by a HIS report and number of matching patients identified by manual expert review. The HIS report queried diagnosis code (I48.11 or I48.0) for the department of cardiology. In this example, all patients listed on the report were analyzed manually, i.e. $s_T = n_T$.

Within seven months (November 2007 to May 2008) 544 patients were found in the HIS report; all these patients were reviewed manually and 304 matching patients were found, i.e. $T = [\text{November 2007; May 2008}]$, $n_T = 544$, $m_T = 304$, $s_T = 544$.

When looking at data values of Table 1, it is striking that the number of matching patients is very low in March, April and May 2008. Pearson's Chi-squared test to compare proportions of $m_T/(s_T - m_T)$ by t results in a highly significant p -value ($p < 2.2E-16$), therefore our estimation formula 1 cannot be applied in this example.

Example 2: Leukemia Trial

In analogy to example 1, ► Table 2 presents number of patients identified by HIS report-1 and associated number of matching patients. This report queried diagnosis code (C92.0-, C92.00 or C92.01) for the department of oncology. Again, all patients were analyzed manually. Within six months (April 2008 to September 2008) 283 patients were listed in HIS report-1. Twenty-eight match-

Table 1 A HIS report generates monthly lists of potential trial patients for an atrial fibrillation trial (second column). Experts reviewed manually medical records from these persons and identified matching patients for the trial (third column). Overall, 304 of 544 (56%) of HIS report patients were suitable for the trial.

| month | number of patients in HIS report per month ($n_t = s_t$) | number of matching patients from manual expert review per month (m_t) |
|---------------|--|---|
| November 2007 | 79 | 71 |
| December 2007 | 60 | 55 |
| January 2008 | 76 | 62 |
| February 2008 | 90 | 71 |
| March 2008 | 70 | 21 |
| April 2008 | 96 | 21 |
| May 2008 | 73 | 3 |
| total | $n_T = s_T = 544$ | $m_T = 304$ |

ing patients were identified by manual review, i.e. $T = [\text{April 2008}; \text{September 2008}]$, $n_T = 283$, $m_T = 28$, $s_T = 283$.

Pearson's Chi-squared test to compare proportions of $m_T/(s_T - m_T)$ by t results in a non-significant p -value ($p = 0.60$), therefore our estimation formula 1 can be applied.

Formula 1 yields an estimated patient accrual rate $PAR = 4.67/\text{month}$ with a 95% confidence interval (3.15/month; 6.59/month).

When comparing Table 1 and Table 2 it is striking that the overall proportion of matching patients is much lower in Table 2. Therefore we applied an improved HIS report-2 which eliminated persons with previous leukemia episodes as well as duplicate persons (► Table 3). With this improved report, n_T was reduced ($n_T = s_T = 53$) for the same number of matching patients ($m_T = 28$), i.e. 53% of HIS report-2 patients were suitable for the trial. Again, Pearson's Chi-squared test to compare proportions of $m_T/(s_T - m_T)$ by t results in a non-significant p -value ($p = 0.13$), therefore our estimation formula 1 can be applied. Formula 1 yields an estimated patient accrual rate $PAR = 4.67/\text{month}$ with a 95% confidence interval (3.41/month; 5.89/month).

Discussion

In Germany and many other countries, electronic HIS are available in almost all hospitals. Initially, they were implemented for administrative purposes (billing, DRG system), but in recent years more and more clinical information is available in these systems. Due to deficiencies in data monitoring and software validation they are at present not suited for documentation of clinical trials, but they contain relevant information, such as diagnosis codes, which can be used to support patient recruitment [4].

Estimation of realistic patient accrual rates is important for planning of clinical trials, but quite difficult. The phenomenon that patient recruitment often takes much more time than investigators expected is called "Lasagna's Law" [10] (Louis Lasagna, clinical pharmacologist, investigator of the placebo response). Collins [11] wrote about "fantasy and reality" of patient recruitment and concluded "we cannot overemphasize the importance of paying adequate attention to

Table 2 Leukemia trial. HIS report-1 selects potential trial patients based on ICD codes (second column). Matching patients were identified by manual review of medical records (third column). Overall, only 28 of 283 (9.9%) of HIS report-1 patients were suitable for the trial.

| month | number of patients in HIS report-1 per month ($n_t = s_t$) | number of matching patients from manual expert review per month (m_t) |
|----------------|--|---|
| April 2008 | 49 | 2 |
| Mai 2008 | 30 | 5 |
| June 2008 | 52 | 5 |
| July 2008 | 63 | 6 |
| August 2008 | 47 | 5 |
| September 2008 | 42 | 5 |
| total | $n_T = s_T = 283$ | $m_T = 28$ |

Table 3 Leukemia trial. In contrast to Table 2, HIS report-2 eliminates persons with previous leukemia episodes as well as duplicate persons (second column). Matching patients were identified by manual review of medical records (third column, same as in Table 2). Overall, 28 of 53 (53%) of HIS report-2 patients were suitable for the trial.

| month | number of patients in HIS report-2 per month ($n_t = s_t$) | number of matching patients from manual expert review per month (m_t) |
|----------------|--|---|
| April 2008 | 6 | 2 |
| Mai 2008 | 10 | 5 |
| June 2008 | 13 | 5 |
| July 2008 | 6 | 6 |
| August 2008 | 11 | 5 |
| September 2008 | 7 | 5 |
| total | $n_T = s_T = 53$ | $m_T = 28$ |

sample size calculations and patient recruitment during the planning process. A sample size that is too small may turn a potentially important study into one that is indecisive or even an utter failure". There is a lot of evidence that many clinical trials failed behind their recruitment objectives [1, 2]. Data monitoring committees must frequently decide about actions in trials with lower-than-expected accrual [12]. Carter [13] stated "the most complicated aspect pertaining to the estimation of accrual periods is the determination of the expected rate".

HIS statistics can be used to estimate annual case numbers for a specific disease. However, this approach lacks precision, because due to specific inclusion and exclusion criteria only a subset of these patients is

eligible for a certain trial. Depending on these criteria, the rate of suitable patients within a certain disease may vary considerably. For this reason we combine a HIS report with manual expert review of patient records to estimate possible accrual rates more precisely. Manual chart review is labor-intensive; especially when n_T is large, analysis of a sample s_T ($s_T \ll n_T$) may provide an acceptable estimation of accrual rate.

Our second example (► Table 2) demonstrates that simple HIS statistics like annual case numbers for a certain disease can substantially overestimate patient accrual rates. HIS are case-centric, not patient-centric: For example, an AML patient with six chemotherapy cycles may be represented in six HIS cases. Therefore it is necessary to design HIS

reports that aggregate information on several cases by patient. Appropriate HIS reports for clinical data retrieval are non-trivial, in particular if temporal relations are taken into account [14]. ▶ Table 3 presents output from a more complex HIS report, which provides more accurate data.

Patient lists from HIS reports can be pseudonymized easily, for instance using reference numbers instead of patient names. In contrast, manual expert review of individual patient records implicates access to identifiable data, because these patient charts contain a large proportion of unstructured text elements (e.g. physician letters). Therefore data protection rules need to be applied. According to German law, a physician at a university hospital who is involved in care for a specific patient, is allowed to analyze this data for scientific purposes. In general, the data access policy for patient data needs to be approved by the responsible data protection officer.

Our approach does not take into account that only a subset of eligible patients decide to participate in a trial. This rate also depends on many factors, which are difficult to be quantified, such as assessment of risks and benefits, motivation and beliefs of physician, as well as organizational infrastructure of a specific trial. Our method depends on quality of HIS data. Complete and correct HIS data is needed for precise accrual rate estimations. Diagnosis codes play a key role for inclusion and exclusion of patients in clinical trials. The same data is very relevant for billing purposes in a DRG system, therefore – at least in Germany – these data items are monitored intensively by physicians, hospital administration and health insurances.

For various reasons – for instance change of healthcare service structures or disease incidence – patient accrual rates can change over time. Therefore we propose to assess proportions of matching and non-matching patients over time by means of Pearson's Chi-

squared test. Our first dataset was not appropriate for patient accrual rate estimation because of outliers, while our second example was suitable for this procedure.

Secondary use of routine HIS data for scientific purposes becomes attractive, because systems get mature and the amount of available data is growing over time. A recent study reports that a large proportion of data for trials can be derived from routine data [15], leading to the visionary concept of “single source”, i.e. using HIS data directly for clinical trials [16].

Conclusion

HIS-based estimation of patient accrual rates is feasible and should be applied to improve planning of clinical trials.

References

1. Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. *Br Med J (Clin Res Ed)* 1984; 289 (6454): 1281–1284.
2. Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, Entwistle V, Garcia J, Roberts I, Grant A. Recruitment to randomised trials: strategies for trial enrolment and participation study. The STEPS study. *Health Technol Assess* 2007; 11 (48).
3. Mapstone J, Elbourne D, Roberts. Strategies to improve recruitment to research studies (Review). *Cochrane Database Syst Rev* 2007; (2): MR000013.
4. Dugas M, Lange M, Berdel WE, Müller-Tidow C. Workflow to improve patient recruitment for clinical trials within hospital information systems – a case-study. *Trials* 2008; 9: 2.
5. R: A language and environment for statistical computing. <http://www.R-project.org>.
6. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; 26: 404–413.
7. Kirchhof P, Auricchio A, Bax J, Crijns H, Camm J, Diener HC, Goette A, Hindricks G, Hohnloser S, Kappenberger L, Kuck KH, Lip GY, Olsson B, Meinertz T, Priori S, Ravens U, Steinbeck G, Svernhage E, Tijssen J, Vincent A, Breithardt G. Outcome parameters for trials in atrial fibrillation: executive summary. *Eur Heart J* 2007; 28 (22): 2803–2817.
8. Büchner T, Hiddemann W, Berdel WE, Wörmann B, Schoch C, Fonatsch C, Löffler H, Haferlach T, Ludwig WD, Maschmeyer G, Staib P, Aul C, Grun-eisen A, Lengfelder E, Frickhofen N, Kern W, Serve HL, Mesters RM, Sauerland MC, Heinecke A; German AML Cooperative Group. 6-Thioguanine, cytarabine, and daunorubicin (TAD) and high-dose cytarabine and mitoxantrone (HAM) for induction, TAD for consolidation, and either prolonged maintenance by reduced monthly TAD or TAD-HAM-TAD and one course of intensive consolidation by sequential HAM in adult patients at all ages with de novo acute myeloid leukemia (AML): a randomized trial of the German AML Cooperative Group. *J Clin Oncol* 2003; 21 (24): 4496–4504.
9. Büchner T, Berdel WE, Schoch C, Haferlach T, Serve HL, Kienast J, Schnittger S, Kern W, Tchinda J, Reichle A, Lengfelder E, Staib P, Ludwig WD, Aul C, Eimermacher H, Balleisen L, Sauerland MC, Heinecke A, Wörmann B, Hiddemann. Double induction containing either two courses or one course of high-dose cytarabine plus mitoxantrone and post-remission therapy by either autologous stem-cell transplantation or by prolonged maintenance for acute myeloid leukemia. *J Clin Oncol* 2006; 24 (16): 2480–2489.
10. Lasagna L. Problems in publication of clinical trial methodology. *Clin Pharmacol Ther* 1979; 25 (5 Pt 2): 751–753.
11. Collins JE, Williford WO, Weiss DG, Bingham SE, Klett C. Planning patient recruitment: fantasy and reality. *Stat Med* 1984; 3 (4): 435–443.
12. Korn EL, Simon R. Data monitoring committees and problems of lower-than-expected accrual or events rates. *Control Clin Trials* 1996; 17 (6): 526–535.
13. Carter RE, Sonne SC, Brady KT. Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC Medical Research Methodology* 2005; 5: 11.
14. Dorda W, Gall W, Duftschmid G. Clinical data retrieval: 25 years of temporal query management at the University of Vienna Medical School. *Methods Inf Med* 2002; 41 (2): 89–97.
15. Williams JG, Cheung WY, Cohen DR, Hutchings HA, Longo MF, Russell IT. Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment. *Health Technol Assess* 2003; 7 (26): iii, v-x, 1–117.
16. Kush R, Alschuler L, Ruggeri R, Cassells S, Gupta N, Bain L, Claise K, Shah M, Nahm M. Implementing Single Source: the STARBRITE proof-of-concept study. *J Am Med Inform Assoc* 2007; 14 (5): 662–673.