Personal medical data linking: Development and validation of a reliable and easy-to-use software tool

Journal:	*Methods of Information in Medicine*
Manuscript ID	Draft
Manuscript Type:	Original Article for a Focus Theme
Date Submitted by the Author:	n/a
Complete List of Authors:	orazio, Sébastien; Haematological Cancer Registry of Gironde; Inserm Unit U1219, EPICENE team, University of Bordeaux Maurisset, Sylvain; Gironde General Cancer Registry Degre, Delphine; Manche General Cancer Registry, Centre Hospitalier Public du Cotentin Billon-Delacour, Solenne; Loire-Atlantique and Vendée General Cancer Registry Poncet, Florence; Isère General Cancer Registry Colonna, Marc; Isère General Cancer Registry Monnereau, Alain; Haematological Cancer Registry of Gironde; Inserm Unit U1219, EPICEN team, University of Bordeaux
Keywords:	Record linkage, software, cancer registry, computer program, identity matching

----

Personal medical data linking: Development and validation of a reliable and easy-to-use software tool.

S.Orazio - Haematological Cancer Registry of Gironde Bordeaux, France and Inserm Unit U1219, EPICENE team, University of Bordeaux, France.

S. Maurisset - Gironde General Cancer Registry, Bordeaux, France.

D. Degre - Manche General Cancer Registry, Centre Hospitalier Public du Cotentin, Cherbourg, France.

S. Billon-Delacour - Loire-Atlantique and Vendée General Cancer Registry, Nantes, France.

F. Poncet - Isère General Cancer Registry, Grenoble, France.

M. Colonna - Isère General Cancer Registry, Grenoble, France.

A. Monnereau - Haematological Cancer Registry of Gironde, Bordeaux, France and Inserm Unit U1219, EPICENE Team, University of Bordeaux, France.

Correspondance to:
Sébastien Orazio
Registre des hémopathies malignes de la Gironde
Institut Bergonié
229 cours de l'Argonne
33076 Bordeaux Cedex
France
Telephone : +33556330484
Fax : +33547306072
E-mail : s.orazio@bordeaux.unicancer.fr

Summary

Objectives

To propose a reliable and easy-to-use tool to link medical databases based on the latest scientific advances in bioinformatics and biostatistics. A semi-automatic linking tool has to provide a list of possible pairs, while optimising the cost (in terms of amount of manual verifications) / effectiveness (in terms of recall and precision of the system) ratio depending on user priorities.

Methods

We developed a package with the R software including the main steps to link two databases: 1- cleaning and data standardizations, 2- management of multiple names and surname or patronymic name, 3- a mixed of deterministic and probabilistic record linkage, 4- output files return a list of linkage. We used the *P. Contiero* probabilistic approach to product global weights in order to distinguish matches from non-matches. For more flexibility, we computed acceptability threshold by unsupervised procedure based on extreme value statistics (EVT) concepts.

Efficiency of our algorithm is evaluated on real data by the cost/efficacy ratio, with the cost defined by the number of manual verifications and efficacy measured with the F-measure indicator.

Results

The F-measure result of our algorithm was 0.99 for a mean computation time of 58s on the evaluation dataset (3,535 x 39,660 identities). The number of manual validations was 188 pairs (5.3% of the source file).

Conclusion

The algorithm is portable, flexible and efficiency. Calibrated with a dataset of a medium size from the French cancer registries, our algorithm can be adapted (by new R-language program lines) to bigger databases or other structured data in order to yield powerful results. However, further evaluations are needed to take into account other kinds of empirical or artificial data.

Keywords

Record linkage, software, cancer registry, computer program, identity matching

----

1.Introduction

Record linkage refers to the process in which records referring the same entities are detected in different databases or in a unique database (data deduplication). One typical application of record linkage is the collection of different medical datasets in a cancer registry and the deduplication of patient identities in order to avoid overestimation of cancer incidence.

In France, the cancer surveillance is based on the French cancer registries' network (Francim). These registries record all new cancer cases continuously and cover approximately 20% of the national territory. To achieve exhaustiveness, our process uses an active search of incident cancer cases by linking personal medical data from all available information sources (French Hospital Discharge Data System [PMSI], clinical and pathological laboratories, cancer networks, Hospital registries [EPC]). Linking different files from various sources is the core function of the registries and is also commonly used in cancer research. Especially in France, researchers do not have access to the unique identifier of patients, as may be the case in other countries. The choice of the record linkage tool is crucial because the inclusion of cases in the registry depends on it [1].

On the other hand, the high contribution of registries to epidemiological research has necessitated the development of new linkage tools, particularly with cohorts. Again, the choice of linkage tool is very important for analyses such as estimation of survival or incidence [2, 3].

A quick review of the 26 French cancer registries has shown heterogeneity among the techniques used. Most registries used a deterministic technique but some registries calculated an overall score to determine whether two identities (possible pairs) are really identical or not. However, these registries didn't use a probabilistic algorithm to calculate this score.

From a methodological point of view, linking source data is a particularly complex problem that, since Fellegi & Sunters's first studies in 1969 [4], has established itself as an actual scientific research field and most publications on the topic demonstrate the power of probabilistic methods [5,6]. In spite of this, and to date, no consensus has been reached on any algorithm. The most evolved techniques, be it commercial or free software solutions, are costly or difficult to install and use.

A reliable, portable and easy-to-use linkage tool for personal medical data is particularly relevant given the increase the number of electronic databases and the need to identify cancer patients from these databases.

2.Objectives

Our main objective was to propose a reliable and easy-to-use tool to link two medical databases based on the latest scientific advances in bioinformatics and biostatistics. The linking tool has to be semi-automatic with a list of possible pairs being provided, that optimises the cost (in terms of amount of manual verification) / effectiveness (in terms of recall and precision of the system) ratio depending user priorities.

3.Methods

3.1 Choice of technology

We choose to develop our algorithm in the S+ language in order to propose a package easy-to-use with the R software. This allows us to use the already existing tools of the Record Linkage package developed by M. Sariyar and A. Borg [7].

3.2 Overview

We propose to manage together the steps involved in a data linkage strategy: 1/ preparation of the data, 2/ deterministic and probabilistic linkage, 3/ verification of the possible pairs. As we usually do in our identity management programme in cancer registries, the identification elements taken into consideration included patronymic and marital names, surname, birthdate and place of residence (postcode). We used a combination of deterministic and probabilistic approaches in order to take advantage of each technique. We calibrated the algorithm by linking identity data of cancer cases (from 2002 to 2013) from the Gironde haematological cancer registry (source: 10,032 identities) and the Gironde general cancer registry (target: 86,794 identities). The efficacy was evaluated after verification of the identity pairs resulting from the linkage of a multidisciplinary team meetings file (MDT file from our regional cancer network "Réseau de Cancérologie d'Aquitaine"; source: 3,535 identities) and the references identities dataset from the Gironde haematological cancer registry (from 2002 to 2013, target: 39,660 identities). In France at least one MDT is indicated for all new cancer cases. The MDT file record information on all patients that have been discussed for therapeutic decision. No specific identities checks are routinely applied when recording the MDT data. This situation put us in the worst situation regarding this poor identities quality.

3.3 Data pre-processing

The first step consists of specific data cleaning techniques and data standardizations. We removed special characters (punctuation, comas, accents etc.), useless spaces and we have used only uppercase [8]. We have added a step for the management of multiple names, surnames or patronymic names: we duplicated one patient's line for all possible combinations of multiple marital names, patronymic names and surnames (figure 1).

3.4 Deterministic approach

We proposed to add deterministic approximation steps to reduce the number of pairs needed to be integrated in a probabilistic search. Indeed, the probabilistic approach needs to treat information on all the possible pairs, which is n x m possible pairs (more than a billion pairs in the calibration files) with the risk of taking too much space on the PC RAM. By completely removing identical pairs on "name + surname + birth date", we also reduced the use of computing resources. This deterministic approach was computed in 1st and 2nd positions, i.e. before and after the management of multiple names and surnames.

3.5 Stochastic approach

The classical record linkage framework is based on probabilistic models and was outlined by Fellegi and Sunter [6]. This model used conditional probabilities to compute weights of the form:

$$W_{\gamma} = \log(\frac{P(\gamma|Z = 1)}{P(\gamma|Z = 0)})$$

These weights called global weights are used in order to discern matches and non-matches.

If only weights are to be computed without relying on the assumptions of probabilities, then simple methods like the one implemented by *P. Contiero* are suitable [7,9] and that is the case here. Indeed, we didn't want to discern matches and non-matches, but we intended to propose a listing of patient with optimal probabilities of linkage. To this end, we use the "EpiWeigths" function available in the RecordLinkage R package. It is a simple and straightforward procedure within the scope of the Fellegi-Sunter model. In this way, the general formula for comparing the records is:

$$S(\underline{X}, \underline{t}) = \frac{\sum_i w_i s(\underline{X_i}/\underline{t_i})}{\sum_i w_i}$$

$S(\underline{X}, \underline{t})$ is the global weight for same data pairs and is calculated for each record from the source. w_i is the weighting assigned to the ith field. w_i is constituted by the error rate, e_j and average frequency of values in the field, f_j. Average frequency f_j has to be estimated using available data. Error rate e_j depends of the fields chosen for linkage. Following the suggestions by *P. Contiero [9],* we propose the following error rates: name 0.05, surname 0.02, date of birth 0.03 and postcode 0.01.

3.6 Application of string metrics

The record linkage could be assimilated to an extension of a string identification task when errors occur. In this perspective, we used string metrics to adjust the corresponding individual weights for exact agreement. Important string metric algorithms are N-grams, edit-distance (Levenshtein) and Jaro-Winkler string metric procedures. Some empirical studies have shown that differences in the mentioned string metrics are negligible [10]. In our method, the string metrics established by Levenshtein were used because the computing time seemed faster.

3.7 Acceptability threshold

The "EpiWeigths" function only identifies similarities between the two records under comparison. So, the user must impose a threshold for the corresponding percentages between the source and target records used in the linking. We chose to compute acceptability threshold by unsupervised procedure based on extreme value statistics (EVT) concepts. A mean residual life plot ("getParetoThreshold" R function) is generated on which the interval (I-EVT) representing the relevant area for false match rates is to be determined [7, 11]. Based on the assumption that this interval corresponds to a fat tail of the empirical weights distribution, the generalized Pareto distribution is used to compute the acceptability threshold.

3.8 Blocking fields

Blocking is a common strategy to reduce computation time and memory consumption by only comparing records with equal values for a subset of attributes, called blocking fields [7, 10]. This step is important because, in the R software, version 3.2.3, vector size is limited (~250mb). We chose the block with a 2x2 matrices for all restricted comparison between name, surname, postcode, birth day, birth month and birth year. The cutoff value is defined by:

$$c = \frac{lim_p}{Total_p} \times 0.1$$

With lim_p the limited vector size in R, and $Total_p$ the maximal vector size (with calibrating dataset) in the case of unrestricted comparison patterns.

3.9 Efficiency evaluation

Efficiency of our algorithm is evaluated by the cost/efficacy ratio, with the cost defined by the number of manual verifications and efficacy measured with the F-measure indicator [12, 14].

4.Results

4.1 Calibrating the algorithm

Following the first steps (cleaning, standardizations, management of multiple names/surnames/patronymic names, and deterministic record linkage), 1.7 billion pairs were evaluable with the calibration dataset (10,032 x 86,794 identities plus their names decompositions). We choose blockings that were < 0.0013 from table 1.

In the calibration dataset, we noticed missing data for postcode (4.5% in calibration dataset and 14% in the evaluation dataset). Hence, we added nine probabilistic linkages (see table 2, compute positions from 16 to 24) without a postcode field.

Each blocking fields were computed in the order shown in the table 2. Blocking fields were classified from the smallest to the largest *c* index.

Since the first results have sometimes shown complete errors in name or surname or birth date while other fields corresponded perfectly, we added three deterministic linkages: 25- name + surname + postcode, 26-name + birth date + postcode, 27-surname + birth date + postcode.

4.2 Return list of linkage

Table 3 gives an example of return list of linkage. Compute position from number 1-2 and 25-27 correspond to the deterministic record linkage. Others correspond to the probabilistic record linkage with appropriate blocking fields. In this example (table 3), optimal computed threshold was 0.844 for the compute position 8 (Block “name, birth month”). Link is T (True) when only one corresponded with the deterministic algorithm. Link with P (Probable) required manual validation. This return list is the input on “.csv” files.

4.3 Efficiency

The F-measure of our algorithm was 0.99 for a mean computation time of 58s, on the evaluation dataset. The number of manual validations was 188 pairs (5.3% of the source file). We varied the I-EVT (80-140% I-EVT) to propose three moderate settings of our algorithm: cheaper (1.3), optimized (1), sensitive (0.80) (Figure 2).

5.Discussion and Conclusion

Our aim was to develop a record linkage system that was easy-to-use, portable, and that integrated sophisticated linking processes.

The different steps developed were computed in a R package[1]. The ease-of-use was achieved by this simple R function that allowed the user freedom to impose I-EVT without the need to modify the source code. The user can easily choose the cost/efficacy ratio in agreement with results of our validation tests.

The algorithm is portable because R can be installed on Windows-based or Linux-based PCs. The algorithm is flexible and the threshold itself can be adapted to the data. Indeed when applying EVT, we do not need training data or other supervised technique for the determination of a threshold [15].

The algorithm is efficient and the output files on ".csv" format allow simple integration of results in another system. Information on weights, type of research/block and True or Probable links, improved facilities for clerical review and highlighting of agreement or disagreement in records pairs.

During its development, the package functions were used in the Gironde Hematological and General Cancer Registries for a period of one year, with good results. However, our algorithm was calibrated with data from French cancer registries and using it with differently structured data could yield less powerful results [12]. Furthermore, other evaluations on very different sorts of empirical or artificial data are needed.

Also, our treatment of unknown comparison values was probably too trivial. More sophisticated approaches existing to dealing with missing values could be added, but their results need to be evaluated [16].

On the other hand, the nature of similarity or stochastic functions used has an important influence on linkage efficiency [13]. In particular, the stochastic record linkage based on the specific EM algorithm seem to produce the best (~1% more) classification results when calibrating data are structurally different to validation data [12, 17]. Therefore, the EM algorithm could be a good alternative when our method is applied on differently structured data compared those of the French cancer registries. Furthermore, we designed our package such that, in future versions, the similarity (added Jaro-Winkler) and the stochastic (added EM

[1] You can upload the package to http://etudes.isped.u-bordeaux2.fr/registres-cancers-aquitaine/General/document/concordantSearch_0.9.1.zip

algorithm) functions can be changed or redefined by the user. Whether or not this is necessary will depend on the results of the ongoing tests using other databases.

Acknowledgments

The authors would like to thank Dr. Ravi Nookala and Dr. Jone Iriondo-Alberdi of Institut Bergonié for the medical writing services and to Dr. Brice Amadeo for the review.

References

1-Clark DE. Pratical introduction to record linkage for injury research. Inj Prev 2004; 186-191.

2-Oberaigner W. Errors in Survival Rates Caused by Routinely used deterministic record linkage methods. Methods Inf Med 2007; 420-424.

3-Moor CL,Gidding HF, Law MG, Amin J. Poor record linkage sensitivity biased outcomes in a linked cohort analysis. Journal of Clinical Epidemiology; 2016; 16: 61-65.

4-Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association 1969; 1183-1210.

5-Silveira DP, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systemic review. Rev Saude Publica 2009; 43-45.

6-Boyd JH, et al. Technical challenges of providing record linkage services for research. BMC Med Inform Decis Mak 2014; 14-23.

7- Sariyar M and Borg A. The RecordLinkage Package: Detecting Errors in data. The R Journal 2010; 2: 61-67.

8-Churches T, Christen P, Lim K and Zhu J. Preparation of name and address data for record linkage using hidden Markov models. BMC Med Inform Decis Mak 2002; 2-9.

9-Conteiro P, Tittarelli A, Tagliabue G, Maghini A, Fabiano S, Crosignani P, Tessandori R. The Epilink Record Linkage Softaware. Methods Inf Med 2005; 44: 66-71.

10-Christen P, Goiser K. Quality and complexity measures for data linkage and deduplication. In F. Guillet and H. Hamilton, editors, Quality Measures in Data Mining, Studies in Computational Intelligence. Springer, 2006.

11-Sariyar M, Borg A, Pommerening K. Controlling false match rates in record linkage using extreme value theory 2011; 44: 648-654.

12-Sariyar M, Borg A, Pommerening K. Evaluation of Record Linkage Methods for Iterative Insertions. Methods Inf Med 2009; 48:429-437.

13-Belin T, Rubin D. A method for calibrating false-match rates in record linkage. J Am Stat Assoc 1995;90: 694-707.

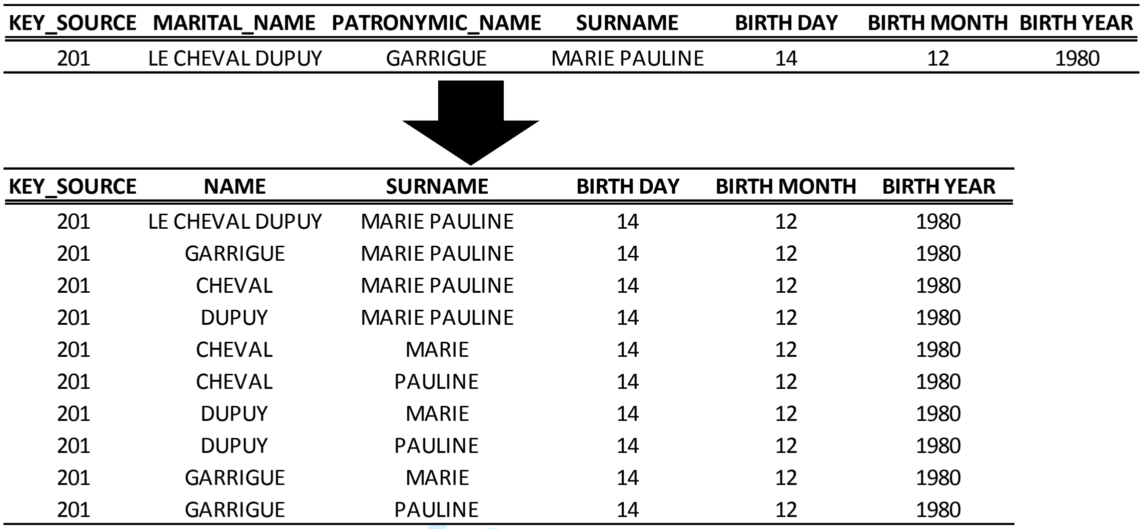
14-Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol 2002;31: 1246-1252.

15-Sariyar M, Borg A, Pommerening K. Active learning strategies for the deduplication of electronic patient data using classification trees. J Biomed Inform 2012;45: 893-900.

16-Sariyar M, Borg A, Pommerening K. Missing values in deduplication of electronic patient data. J Am Med Inform Assoc 2012;19: e76-e82.

17-Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. Am Med Inform Assoc Symposium Proceeding 2003; 259-263.

Figures and Tables



KEY_SOURCE	MARITAL_NAME	PATRONYMIC_NAME	SURNAME	BIRTH DAY	BIRTH MONTH	BIRTH YEAR
201	LE CHEVAL DUPUY	GARRIGUE	MARIE PAULINE	14	12	1980

KEY_SOURCE	NAME	SURNAME	BIRTH DAY	BIRTH MONTH	BIRTH YEAR
201	LE CHEVAL DUPUY	MARIE PAULINE	14	12	1980
201	GARRIGUE	MARIE PAULINE	14	12	1980
201	CHEVAL	MARIE PAULINE	14	12	1980
201	DUPUY	MARIE PAULINE	14	12	1980
201	CHEVAL	MARIE	14	12	1980
201	CHEVAL	PAULINE	14	12	1980
201	DUPUY	MARIE	14	12	1980
201	DUPUY	PAULINE	14	12	1980
201	GARRIGUE	MARIE	14	12	1980
201	GARRIGUE	PAULINE	14	12	1980

Figure 1. Management of multiple names and surnames or patronymic names. Identities are not true patients.

Table 1. "c" value with calibration dataset (1.7 billion of pairs possible). "c" value > 0.0013 are in grey color.

Blocking	1-name	2-surname	3-Postcode	4-birth day	5-birth month	6-birth year
1-name	0,000187	0,000002	0,000006	0,000006	0,000016	0,000004
2-surname	0,000002	0,009935	0,000212	0,000324	0,00083	0,000235
3-Postcode	0,000006	0,000212	0,022009	0,000713	0,001838	0,000386
4-birth day	0,000006	0,000324	0,000713	0,002437	0,002753	0,000580
5-birth month	0,000016	0,00083	0,001838	0,002753	0,083433	0,001491
6-birth year	0,000004	0,000235	0,000386	0,00058	0,001491	0,017816

Table 2. Blocking fields, fields for probabilistic linkage and corresponding I-EVT.

Compute position	Blocking fields	fields for probabilistic Linkage	I-EVT
Deterministic approach			-
1	all	-	-
2	name, surname, birth date	-	-
Stochastic approach			
3	name, surname	postcode, birth date	[0,51;0,81]
4	name, birth year	surname, postcode, birth date	[0,72;0,88]
5	name, postcode	surname, birth date	[0,64;0,82]
6	surname, postcode	name, birth date	[0,64;0,82]
7	name, birth day	surname, postcode, birth date	[0,72;0,90]
8	name, birth month	surname, postcode, birth date	[0,78;0,88]
9	name only	surname, postcode, birth date	[0,72;0,84]
10	surname, birth year	name, postcode, birth date	[0,74;0,84]
11	surname, birth day	name, postcode, birth date	[0,72;0,84]
12	postcode, birth year	name, surname, birth date	[0,71;0,82]
13	birth day, birth year	name, surname, postcode, birth date	[0,71;0,95]
14	postcode, birth day	name, surname, birth date	[0,71;0,94]
15	surname, birth month	name, postcode, birth date	[0,72;0,82]
16	name, surname	birth date	[0,71;0,88]
17	name, birth year	surname, birth date	[0,72;0,86]
18	name, birth day	surname, birth date	[0,70;0,92]
19	name, birth month	surname, birth date	[0,64;0,82]
20	name only	surname, birth date	[0,74;0,90]
21	surname, birth year	name, birth date	[0,70;0,88]
22	surname, birth day	name, birth date	[0,68;0,82]
23	birth day, birth year	name, surname, birth date	[0,70;0,80]
24	surname, birth month	name, birth date	[0,68;0,85]
Deterministic approach			
25	name, surname, postcode	-	-
26	name, birth date, postcode	-	-
27	surname, birth date, postcode	-	-

Table 3. Examples of return list of linkage. Identities are not true patients

KEY	MARITAL NAME - SURNAME - PATRONYMIC NAME - POSTCODE - BIRTH DATE	FILE	WEIGTHS	COMPUTE POSITION	LINK
9350	ORAZIO - JEANNE MARIE ESTREM - MONJOUST - 33200 - 15/06/1940	SOURCE	1	1	T
9342	ORAZIO - JEANNE MARIE ESTREM - MONJOUST - 33200 - 15/06/1940	TARGET			
119830	BOUZID - BERNADETTE - NA - 33600 - 24/05/1950	SOURCE	1	2	P
44606	BOUZID - BERNADETTE - RACHOU - 33600 - 24/05/1950	TARGET			
119835	RACHOU - BERNADETTE - NA - 33600 - 24/05/1950	SOURCE	1	2	P
44606	BOUZID - BERNADETTE - RACHOU - 33600 - 24/05/1950	TARGET			
23930	MONNEREAU - JEAN - NA - 33180 - 27/02/1979	SOURCE	1	26	P
23929	MONNEREAU - ELPIDIO - MONNEREAU - 33180 - 27/02/1979	TARGET			
120168	LOUVEAU DE LA LEGUYADER - SANDRA - NA - NA - 24/01/1981	SOURCE	0,9575674	8	P
115875	LOUVEAUDE LA LEGUYADER - SANDRE - LOUVEAUDE LAGUIGNERAYE - NA - 24/01/1981	TARGET			

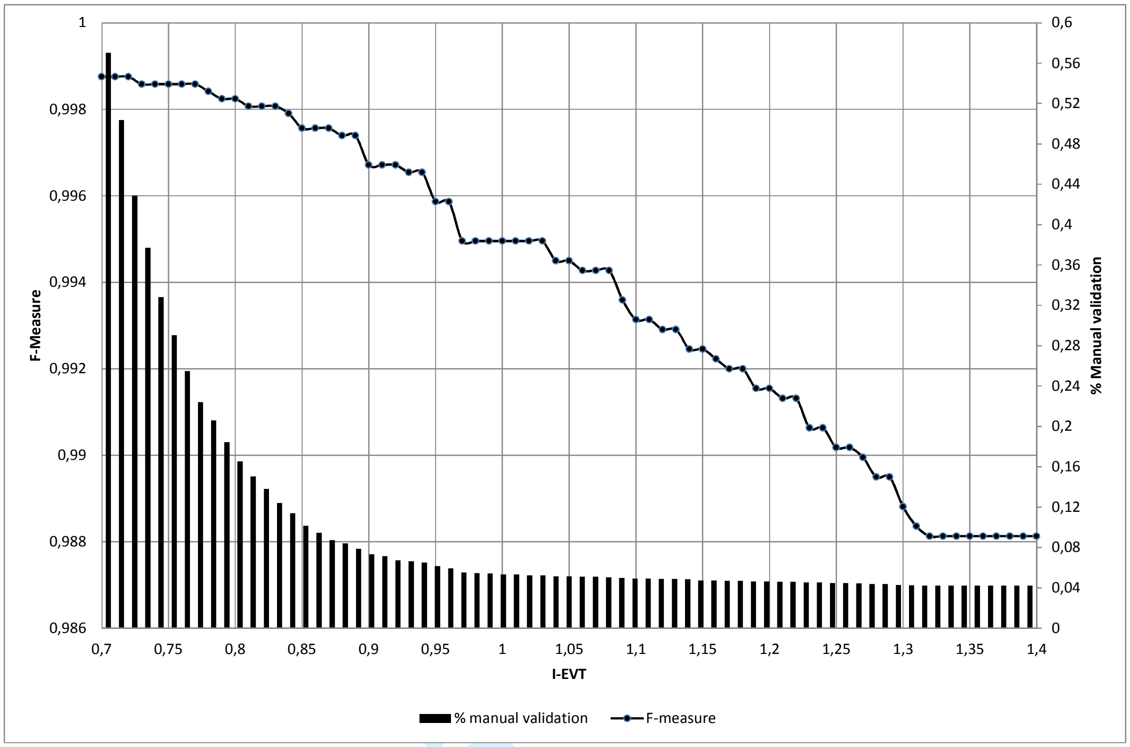


Figure 2. F-measure and number of manual validations / total size of source file for I-EVT varying from 70-140%.