

*Collection of Biostatistics Research Archive*  
COBRA Preprint Series

---

*Year 2006*

*Paper 3*

---

Simple Records Matching Method for  
diagnostic and clinical datasets of patient's  
records

Salvo Reina\*

Vito M. Reina<sup>†</sup>

Eugenio A. Debbia<sup>‡</sup>

\*Institute of Microbiology School of Medicine University of Genoa, Italy, [reina@village.it](mailto:reina@village.it)

<sup>†</sup>ICT Freelance in Rome

<sup>‡</sup>University of Genoa, Italy, [eugenio.debbia@unige.it](mailto:eugenio.debbia@unige.it)

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/cobra/art3>

Copyright ©2006 by the authors.

# Simple Records Matching Method for diagnostic and clinical datasets of patient's records

Salvo Reina, Vito M. Reina, and Eugenio A. Debbia

## Abstract

Several statistical packages, either commercial or open-source, provide many methods for multi-factorial and discriminant analysis; such a software is poorly used by physicians. Appropriate models and tests have to be used pending on the kind of experiment scheme, adequate distribution assumption are needed for variables and parameters and proper data validation have to be verified for historical records. These are but a few of many critical aspects for a robust and trustable data interpretation needed in the Evidence Based Medicine era. Clinicians always wish to be able to quickly interpret diagnostic records to discriminate, or alternatively correlate, coherent groups of patient's records according to either descriptive characters or variable units. Practically, patient's records are stored in spread-sheet or database which change pending on the clinical trial scope; moreover, data entry and its validation is usually poor, hence physician are used to send raw-data to the statistician without contributing, for instance, with parametric and non-parametric indication on usable distribution. We address this problem by introducing a simple "weighted" model approached with the Unique Factorisation Domain theory: records can be compare by matching each other through a score overlap and clinician can modulate tolerance of closeness stringency criteria. An intuitive paradigm of records matching method (RMM) is presented and discussed with example, computational design and programming prototyping model; freely available material concerning real-world application, are also provided by the authors.

## Simple Records Matching Method for diagnostic and clinical datasets of patient's records

Salvo A. Reina<sup>α</sup>, Vito M. Reina<sup>β</sup> and Eugenio A. Debbia<sup>α</sup>

### Abstract

Several statistical packages, either commercial or open-source, provide many methods for multi-factorial and discriminant analysis; such a software is poorly used by physicians. Appropriate models and tests have to be used pending on the kind of experiment scheme, adequate distribution assumption are needed for variables and parameters and proper data validation have to be verified for historical records. These are but a few of many critical aspects for a robust and trustable data interpretation needed in the Evidence Based Medicine era.

Clinicians always wish to be able to quickly interpret diagnostic records to discriminate, or alternatively correlate, coherent groups of patient's records according to either descriptive characters or variable units. Practically, patient's records are stored in spread-sheet or database which change pending on the clinical trial scope; moreover, data entry and its validation is usually poor, hence physician are used to send raw-data to the statistician without contributing, for instance, with parametric and non-parametric indication on usable distribution.

We address this problem by introducing a simple "weighted" model approached with the Unique Factorisation Domain theory: records can be compare by matching each other through a score overlap and clinician can modulate tolerance of closeness stringency criteria.

An intuitive paradigm of records matching method (RMM) is presented and discussed with example, computational design and programming prototyping model; freely available material concerning real-world application, are also provided by the authors.

**Keywords** : records matching, Unique Factorisation Domain, bioinformatics, Evidence Based Medicine

**Corresponding author** : Prof Eugenio A. Debbia ([eugenio.debbia@unige.it](mailto:eugenio.debbia@unige.it))

<sup>α)</sup> *Laboratory of experimental Microbiology and Epidemiology, Dept. DISCAT, School of Medicine, University of Genoa, Italy*

<sup>β)</sup> *Freelance, ICT professional, Rome, Italy*



## 1.0 INTRODUCTION

The method here presented, and its software implementation scheme, is aimed to provide a tool for basic research as well as for clinical data analysis and interpretation. The practical need of a physician is to explore repertoires of clinical dataset to discover similarities in between patient's records as well as to group stratified records according taxonomical and epidemiological criteria.

Mostly, database with clinical data are inferred through Standard Query Language (SQL) with queries submitted to a database server engine. This practise is very effective in grouping coherent *set* of patients according to articulated rank and identity criteria as authors have reported on clinical virology [11]; yet it is not possible to infer with algebraic rules which can discern a tolerance range and smooth clustering criteria.

Several evolute software approaches such as neural network, fuzzy logics and bayesian modelling can be used to treat data [5, 6]. Similarity, proximity and phenotype variability are typical issues related to the every day investigation of Evidence Based Medicine; we felt the need of creating an *apparatus* to treat descriptive correlation across population's data preserving simplicity and easy applicability.

Our method formalises an *apparatus* that can be easily coded with any programming language; thus, scientists can study complex information with records matching method (RMM) rendered with an intuitive software tool. Any set of record profiles can be cross-matched according to multiple-programmable variables or parameters, each of one can be characterised with arbitrary range of weights; plausible ranges of suitable values can be then indicated by the physician's experience without the necessity of statistical assumption.

## 2.0 MODEL DESCRIPTION

Either clinical or biological research provide a panel of scalar as well descriptive factors (either parametric or not parametric); such a panel of information units can be represented as a number called Factorial Record Index (FRI) which renders the entire record as an equivalent numerical value. Multiple experimental data can be then serialised and evaluated for dependencies and proximity according to a set of pre-determined rules. Such a heuristics can be repeated for the fine-tuning and calibration of meta-analysis studies.

Our model allows to assimilate (or discriminate) groups of records according to their affinity, contiguity and closeness even when data sets have diverse numeric sizes.

The model finds which, and estimates how much, a subset of records is similar to a known record considered to be the Master Profile (MP); this MP record can either be a newly inserted or one of the existent record of the analysed table itself. In this latter case the MP is considered against all the other complementary records of the table.

The terms *similarity*, *affinity* as well as for *correlation* and *association* will be meant as appropriate on the base of the meaning of the variables considered. Variables and parameters are practically identified by the fields of a table (or columns of a spreadsheet). Such an homology is implicitly one of the major advantage of the model for its applicability, since can be easily generalised and comprehended by physician with no mathematical expertise.

Any discipline and any kind of descriptive measures in a table (row by columns), can be translate in a range of parametric score as far as a table containing weights-rules (weighted heuristic matrix) has been pre-determined for each unit (column and/or field). By "*summarising*" the field effects a quantitative and qualitative comparison of a couple of records is expressed with a *field-by-field* level of concordancy defined as Matching Level (ML).

The aim of scoring each *record-to-record* ML is concerned with the possibility of sorting and expressing *delta* values so that a simple *cut-off* can separate concordant versus non-concordant records sets.

In the most simple case, the result of a comparisons run is a listing of two sets of records one of which it contains the records which fall in the range of acceptable similarity, whereas the second contains all the others.

## 2.1 MODEL USABILITY IN CLINICS

If coded as a software, the model can be *recursive*, so that the model itself can persist new rules with a back-propagated auto-exploration of the cross-matching results (ML array). This will lead to an auto-calibration of the system which produce an optimal set of rules knowledgebase; this solution can progressively improve RMM competence in scanning historical retrospective databases and cross-drag experimental data evidences coming from different source and periods. Clinical trials and epidemiological meta-analysis fit exactly this expectation.

The software runs reiteration, namely recursive process of the RMM, could also produce self-monitoring logs in order to store typical activity profile; once results are then acknowledged to be effective for a specific discipline a *template* file could provide an objective mean to re-use algorithm and compare data under specific domains with common knowledge rules.

Another stimulating feature of the model's is the *cluster analysis*, especially useful if several fields were trained with *weighted-matrix* on categorical data. Notoriously, either PCA, PCO or *cladistic analysis* are such an heavy task for most of the clinicians, then it would be desirable a much lighter tool.

So far, authors have used RMM on gynecology, antimicrobial therapy, environmental audit measurements and chemical fatty acids mass gas chromatography compounds patterns [17,18].

### 3.0 DEFINITION

3.1) Let us pose

$\mathbf{R} = \{C_1, \dots, C_n\}$  as a record where  $\mathbf{C}$  set of fields for  $\mathbf{n}$  = number of fields;  
*e.g.*  $\mathbf{R} = \{age, job, sex, offspring, marriage\}$  ( $n=5$ );

$C_i = \{v_{i_1}, \dots, v_{i_m}\}$  where  $i$ th is the ordinal of the field in  $\mathbf{R}$ ;  
*e.g.*  $C_4 = \text{offspring} \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$  ( $m=10$ );

$v_{i_j}$  is the value of the  $j$ th ordinal position of the  $i$  field in  $\mathbf{R}$ ; ( $i=1..n$ ;  $j=1..m$ )  
*e.g.*  $v_{4_5} = 5$  (5 children in the fourth field of  $\mathbf{R}$ );

3.2) therefore  $m$  related to  $i$ , is the number of the possible values of the  $i$ th ordinal field  $C_i$  in  $\mathbf{R}$  and it will be noted as  ${}_i\mathbf{M}$ ,

*e.g.*  ${}_4\mathbf{M} = 10$ ; (10 possible values for the fourth field **offspring**);

${}_i\mathbf{Max}$  will represent the maximum among the possible  ${}_i\mathbf{M}$  with  $i=1, \dots, n$ ;

For each  $C_i$  and associated  $i$ , a graph  $G_i$  is defined so that it represents the *weight-distance* between all possible  ${}_i\mathbf{M}$  values contained in  $C_i$ ;

3.3) We define *weight-distance* as an estimation of how much each value  $v_{i_j}$  is unrelated (distant, unlike, different or diverse), to any other  $v_{i_{j^*}}$ ,  $j^* \neq j$  of the field  $C_i$  in  $\mathbf{R}$ . When  $j^* = j$ , thus  $v_{i_j} = v_{i_{j^*}}$ , the *weight-distance* will be zero; as corollary, this implies that when two values are identical their distance will be Null.

The expedient of expressing the concept of *weight-distance* is also worthwhile to measure a quantity of proximity between the field's values; therefore it is determined that a couple of  $(v_{i_j}, v_{i_{j^*}})$  values with  $j^* \neq j$ , will have a *neighbourhood* measurement of the data values for the field to which they belong.

In the example of the **offspring** variable previously cited, the *weight-distance* will be taken as an absolute value obtained as the value's difference or *delta* taken as absolute value. With this simplification the field values coincide with the absolute indexes (ordinal position of the field inside record), albeit this not necessarily has to be the ordinary case.

If we express as  $v_{i_j}$  the **offspring**, being  $v_{i_j} = j$ , it can be considered that two fields are similar (or diverse) as much as the lesser (the higher) is their different **offspring**, thus:

4.4) *weight-distance*  $(v_{i_j} - v_{i_{j^*}}) = |v_{i_j} - v_{i_{j^*}}| = |i - j|$  and if  $v_{44} = 4$  and  $v_{47} = 7$ , *weight-distance*  $= |4 - 7| = 3$

A field will not necessarily contains contiguous values since it can accept various typology and meanings; it also can refer to various scales and applicable values pending on the casting of its validation. More in general, the concept of *weight-distance* will have to be modulated to allow an adequate proximity measurement desired to be consistent with the meaning of the considered information (fields such dates, scores, rank strings, binaries etc.).

#### 4.0 MODEL'S METAPHORE AND REAL-WORLD EXAMPLE

Because we coupled affinity e similarities concepts with geometrical distance values, we initially explain the model by using a geographical metaphor. We use geography of some Italian cities as a scheme of the Record's Matching practically applied and this facilitates theory's approach.

According to point 3.1), let us pose

$$C_4 = \text{City} = \{Rome, Viterbo, Naples, Catania, Milan\}.$$

$C_4$  is the fourth field of a generic record  $R$  which accept 5 descriptive values

It is intuitive and quite plausible to assume the geographical distance between the cities as logical expression of their proximity; this would certainly useful if we would analyse. In this example, that does not make sense to give meaning to the ordinal index of the fields: this is intentionally chosen so no other criteria than kilometer value would indeed mean closeness.

4.1) We start to attribute a weight-arc for each couple of association. All possible arcs produce a graph showed in the Figure 1 which is the visual representation of the matrix formalised in 4.4).

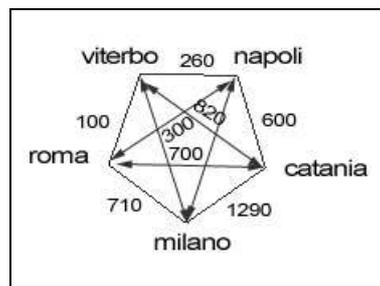


Figure 1 : Geographical graph with Italian cities inter-distances (only main approximate rounded values and original Italian names) For demonstrative purposes the distance reported between nodes are not precise, still the approximation is useful to better understand the formalism, also, the length of each arc in the graph is not proportional for obvious simplicity.

We define  $G_i$  the graph of the *weight-distance* values for each couple of the field in  $R$  and

4.2)  $d_i(j,k)$  for  $j, k=1.., iM$

the *weight-distance* in between the  $j$ th and the  $k$ th ordinal values of the field  $C_i$  in  $R$ .

The  $M_i$  matrix associated with  $G_i$  can be listed as a table which contains the  $d_i(j, k)$  values referred to the field  $C_i$  in  $R$ .

	Rome	Viterbo	Naples	Catania	Milan
Rome	0	100	300	700	710
Viterbo	100	0	260		
Naples	300	260	0	600	
Catania	700		600	0	1290
Milan	710			1290	0

Note that, for this case the matrix  $M_i$  is symmetrical, then  $d_i(j, k) = d_i(k, j)$ ; nevertheless, it is possible to conceive a different definition which lead to a non-symmetrical matrix  $M_i$ , e.g.: a logarithmical  $d(f(j,k))$  function of the

indexed values. We then define  $\underline{M}_i$  a new matrix achieved by normalising  $M_i$  with a pre-determined range of values  $R[0\dots t]$ ; such a range will be common to all the  $\underline{M}_i$ ,  $i=1,\dots,n$ .

## 5.0 THE MATCHING LEVEL OF A RECORD

To compare (determining the similarity) two records, the MP against the Test Profile (TP), all their homologous fields have to be evaluated. Before a practical example few basic assumption have to be postulated.

5.1) Two fields  $C$  and  $C^*$  are comparable only if  $C$  can assume all the possible values of  $C^*$  and *vice versa*.

5.2) If two fields  $C$  and  $C^*$  are comparable, then  $C = C^*$  meaning that they are indeed the same field and share same typology and casting.

5.3) Two records  $R = \{C_1, \dots, C_n\}$  and  $R^* = \{C^*_1, \dots, C^*_m\}$  are comparable only if  $n = m$  and  $C_i$  and  $C^*_j$  are comparable for each  $i=j$ ; for  $i=1,\dots,n$ , and  $j=1,\dots,m$ ; thus

5.4)  $R = \{C_1, \dots, C_n\}$  and  $R^* = \{C_1, \dots, C_n\}$

Evidently, fields as defined in 5.1 can even belong to different tables (foreign dataset) as far as they represents the same meaning for the investigator.

Once we

a) have fixed records  $R = \{C_1(v_j), \dots, C_n(v_n)\}$  and  $R^* = \{C_1(v^*_j), \dots, C_n(v^*_n)\}$  as comparable,

b) have denoted for  $v_j$  and  $v^*_j$  the indexes of the corresponding fields  $C_i$  for  $i=1,\dots,n$ ,

c) have defined  $P_i$  as a parameter associated to the field  $C_i$

we shall define as *Matching Level* or ML of two records the value

$$5.5) \quad K = \sum_{i=1..n} \underline{M}_i(v_j, v^*_j) * P_i$$

The parameter  $P_i$  introduced in 5.5) will be called Weight of Field  $C_i$  or WOF[ $C_i$ ]; such a value indicates the *relevance* of the  $C_i$  field in the overall records matching process. The term “*relevance*” can also be considered as a mean by which it is possible to emphasise, and or reduce, the value of a field; this clearly gives the possibility for a fine-tune of the overall matching calculation and make it possible to qualitatively adjust the contribute of each field to the entire record.

Default value for WOF is  $P_i = 1$  meaning that usually the field accepted value will remain unchanged; yet it could also be possible to selectively exclude a field (variable or column) calculation by assuming  $P_i = 0$  in MP or TP. Noteworthy, WOF [ $P_{ij}$ ] can be used to reduce, rather than amplifying, the relevance of a field; for instance  $P_i = 1/i$ .

## 6.0 INTER-RECORDS AND INTRA-RECORD VARIABLE'S RELEVANCE

In the 5.5) it has been showed how the importance of a field (or variable) can be modulated with the  $P_i$  value; since this is impartially applicable to all the fields, therefore the meaning of the *record-to-record* matching can assume different significance for the *inter-records* comparison. One further level of control on the *cross-matching* process can be implemented at *intra-record* interaction by associating  $P_i$  of two fields on contemporary.

Briefly, under certain circumstances it is useful to consider the concurrency of two variables to either boost or alternatively exclude both  $P_i$  variables contribution.

Also we might want to consider a specific  $P_i$  only for other fields  $C_j$  with  $j \neq i$  so that  $P_i$  depends on the  $i$  and

$$6.1) \quad \mathbf{j} (\mathbf{P}_j \text{ for } \mathbf{j} \neq \mathbf{i})$$

Suppose a data analysis that studies an occupational survey on transporters to measure the prevalence or the incidence of some postural pathologies. In a hypothetical database which contains people profiles with all kinds of works, we should use the Records Matching model by speculating on field **DriverLicence**[A,B,C,E] and field **VehicleType**[*motorcycle, automobile, truck*]

In fact, it is highly probable that when **DriverLicence** =[E] the **VehicleType** =[truck]; this can be translated in our model by “linking” both fields so that  $\mathbf{P}_1$  is function of  $\mathbf{C}_2$ , also:

$$6.2) \quad \mathbf{P}_1 = f(\mathbf{C}_2)$$

Our Records Matching will be conditioned by the relationship  $\mathbf{C}_1 \diamond \mathbf{C}_2$ , and more in general  $\mathbf{P}_i$  linked to  $\mathbf{C}_i$  can be leveraged (enhanced) by the presence of certain value on other fields  $\mathbf{C}_j$  for  $\mathbf{j} \neq \mathbf{i}$ . The appraisal of the set of “enhancers” can be formalised with a three dimensional structure  $\Omega_i$  which contains the values of an enhancer for each index of the values assumed by the field  $\mathbf{C}_i$ .

Essentially, the  $\Omega_i$  represents the structure accounting for the level of exaltation of a enhanced field (variable) on concurrency with the others. The  $\Omega_i$  can be represented as an array containing  ${}_i\mathbf{M} \times \mathbf{n} \times {}_i\mathbf{Max}$  cells or

$$6.2) \quad \Omega_i = \Omega_i ({}_i\mathbf{M}, \mathbf{n}, {}_i\mathbf{Max}).$$

Let us pose  $\Omega_i = \Omega_i ({}_i\mathbf{M}, \mathbf{n}, {}_i\mathbf{Max})$  so that we can define :

$$6.3) \quad \Omega_i (\mathbf{j}, \mathbf{k}, \mathbf{m})$$

as the value which expresses how much the  $\mathbf{m}th$  of the  $\mathbf{k}th$  field can amplify the  $\mathbf{P}_i$  of the field  $\mathbf{C}_i$  when it assumes the value of the  $\mathbf{j}$  index. By common assumption, the  $\mathbf{m}th$  indexed value of the  $\mathbf{k}th$  indexed field has no relationship at all with the WOF[ $\mathbf{P}_i$ ] in the field  $\mathbf{C}_i$  if it assumes the value of the index of  $\mathbf{j}$ , therefore it will be set

$$6.4) \quad \Omega_i (\mathbf{j}, \mathbf{k}, \mathbf{m}) = 1$$

It can also be said that  $\Omega_i$  will contain the value 1 for all those elements belonging to a field which not mutually interfere or interrelate each others.

## 7.0 EXTENDED RECORD'S MATCHING LEVEL

A new variant of the definition of WOF can be achieved by extending the contribute of the field's enhancers at the *record-to-record* matching level. Since value 1 represents the value of a neutral element this will leave unaltered the weight of the corresponding field as well as its influence in the overall record *fingerprint*.

Let us define an extended field's weight according 6.4) :

$$7.4) \quad \underline{\mathbf{P}}_i (\mathbf{i}^*) = \mathbf{P}_i * \prod_{\mathbf{j}=1..n, \mathbf{k}=1..{}_i\mathbf{Max}} \Omega_i (\mathbf{i}^*, \mathbf{j}, \mathbf{k})$$

Where  $\underline{\mathbf{P}}_i (\mathbf{i}^*)$  will be extended weight of the field  $\mathbf{C}_i$  with index  $\mathbf{i}^*$ . Note that the *delta* value  $\mathbf{P}_i$ ,  $\underline{\mathbf{P}}_i (\mathbf{i}^*)$  will depend on the field  $\mathbf{C}_i$  and its index. The value will also be enhanced by the values of the other record's fields  $\mathbf{C}_j$  with  $\mathbf{j} \neq \mathbf{i}$  and this will also impact on the weight definition of ML.

Fixed two comparable records  $\mathbf{R} = \{\mathbf{C}_{1(v_1)}, \dots, \mathbf{C}_{n(v_n)}\}$  and  $\mathbf{R}^* = \{\mathbf{C}_{1(v^*_1)}, \dots, \mathbf{C}_{n(v^*_n)}\}$  and having denoted  $v_i$  and  $v^*_i$  as indexes corresponding to the fields  $\mathbf{C}_i$  for  $\mathbf{i}=1, \dots, \mathbf{n}$ , let

7.5)  $\underline{P}_i, i=1,..n$  the extended weight of the field  $C_i$ ;

Let also associate the field  $C_i$  with the structure  $\Omega_i$  containing the enhancing level of each of the values of the  $C_i$  against each of cross-matched field  $C_j$  for  $j \neq i$ , we shall define Extended Matching Level (ExML) of two records the value:

$$7.6) \quad \underline{K} = \sum_{i=1..n} \underline{M}_i (v_i, v^*_i) * \underline{P}_i (v_i)$$

so that we now describe practical example of the use of the  $\Omega_i$  and other indicators.

Let  $\mathbf{R} = \{\mathbf{Job}(hodman,mechanics,plumber,...), \mathbf{WorkingDays}(1,2,3,4,5,6,7), \mathbf{WorkingHours}(4,8,12),...\}$  and also let  $\underline{\mathbf{R}} = \{(hodman),(5), (4),...\}$  the content a specific record chosen as MP.

One possible inference on the database would group those workers (TP records) who spend an amount of physical stress over a week period of time which can be consider to be similar with  $\underline{\mathbf{R}}$ .

Let now  $\check{\mathbf{R}} = \{(mechanics),(5),(4) \dots\}$  and pose  $\mathbf{d}_1(\mathbf{1}, \mathbf{2}) = \mathbf{k}$ , the *weight-distance* between the first and the second of the possible values of the field  $C_1$  with  $\mathbf{k}$  being a real number. The  $\mathbf{k}$  will express how-much the fields  $C(mechanics)$  and  $C(hodman)$  vary in terms of energy stress for the same period of time. On a physical point of view, it is reasonable to assume that an hodman spends more than a mechanics, thus we can pose  $\underline{P}_i = 0, i \neq 1$ , and  $\underline{P}_1 = 1$ . This states that the field  $C_1$  will be considered “*relevant*” to the record’s matching outcome.

We shall then have:

$$7.7) \quad \underline{K}(\mathbf{R}, \check{\mathbf{R}}) = \mathbf{d}_1(\mathbf{1}, \mathbf{2}) * \underline{P}_1 = \mathbf{k} * 1 = \mathbf{k}$$

which means that the ML between the two records is  $\mathbf{k}$ .

Clearly,  $\underline{K}$  will not be affected by different values assumed by  $C_3$ , yet those values are critical to the matching process in that they represent the working hours per day. We now suppose  $\check{\mathbf{R}}_2 = \{(mechanics),(5), (8), \dots\}$ .

It is evident that the presence of  $C_3(8)$  will affect the  $\underline{K}$  value implying that the hodman and the mechanics have the same stress over a week when the mechanics will have worked double time.

The latter hypothesis can be formalized by altering  $\underline{P}_3$  and  $\underline{M}_3$  but a better choice would be to define an  $\Omega_1$  structure. In fact, we can pose  $\Omega_1(1, 3, 2) = \epsilon$ , for  $\epsilon$  being a real number arbitrary small so that the more the records  $\mathbf{R}$  and  $\check{\mathbf{R}}_2$  will be similar, the smaller  $\epsilon$  it would be. Therefore, considering the example above mentioned, the stress of worker’s category can be equally considered over the week.

Such a situation can be expressed as :

$$7.9) \quad \underline{K}(\mathbf{R}, \check{\mathbf{R}}) = \mathbf{d}_1(1, 2) * \underline{P}_1 = \mathbf{d}_1(1, 2) * \underline{P}_1 * \Omega_1(1,3,2) = \mathbf{k} * 1 * \epsilon = \mathbf{k}\epsilon$$

Since the ExML between records is  $\mathbf{k}\epsilon$ , the presence of  $C_3(8)$  has reduced the *delta* (Euclidian distance) because  $\mathbf{k}\epsilon < \mathbf{k}$  and  $0 \leq \epsilon < 1$ . This is also due to the presence  $\Omega_1(1, 3, 2)$  which explains the similarity of the records.

It appears clear how the value of  $\epsilon$  can be calibrated according investigator’s needs and going back to previous example we can pose  $\Omega_1(1, 3, 2) = \epsilon = 0$  meaning that the two records match completely and their overlap difference is zero.

The case in 7.9) is an extreme simplification for explanatory need, still it can be used for a much complex situation and by consequence the structures  $\underline{M}_i$  and  $\Omega_i$  can leverage in complexity as well.

## 8.0 FACTORIAL RECORD'S INDEX

The identification of a record is important in order to apply the entire RMM and to recursively infer on a database; we shall now provide a unique number which will substitute and identify a record by masking its numeric equivalent. The purpose of an equivalent number for each record is desirable for practical implementation of the method by preventing the analysis of every record fields composition.

We substitute a record profile with its Factorial Record Index (FRI) and because a factorization algorithm will be used, it will also be possible to derive a single FRI going back to each of the field values.

To transform one record in a FRI we have used the theorem of the Unique Factorisation Domain (UFD) and correlated number theories [1, 19, 20]. The preliminary assumption to apply the UFD to our RMM methodology is that all the fields considered for the matching algorithm assume ceased and pre-determined values.

Let defined  $\mathbf{Kn} = (2, 3, 5, 7, 11, \dots, p_n)$  as an array containing all the first  $n$  prime numbers and recall 3.1 formalism

$\mathbf{R} = \{C_1, \dots, C_n\}$  for  $C_i = \{v_{i_1}, \dots, v_{i_{j^*(i)}}, \dots, v_{i_m}\}$ , where  $j^*(i)$  is the index of the value assumed by  $C_i$ ,  $i=1, \dots, n$ , it will be defined as Factorial Record Index or FRI of the record  $\mathbf{R}$ , the following expression :

$$8.1) \quad \Lambda = \mathbf{Kn}(1)^{j^*(1)} * \dots * \mathbf{Kn}(n)^{j^*(n)} = \prod_{i=1..n} \mathbf{Kn}(i)^{j^*(i)}$$

Clearly  $\Lambda$  will be unique for every single record except for the case of identical records. In fact, two FRI  $\Lambda_1$  and  $\Lambda_2$  will be equal if, and only if, all the fields contain the same values. This can practically be the case where a table can have repeated records (identical rows).

The FRI descend from the index of all the fields, therefore it is possible a *reverse process*. Starting from  $\Lambda$  the indexes assumed by the fields will be rescue and associated back the original values contained in a field. To this end it will necessary to create a translation table that will have to be trained with all acceptable and classified values.

## 9.0 MODEL HEURISTICS AND COMPUTATIONAL DESIGN

The process in 8.1) allows to quantitatively calculate a unique record profile, also we pointed out how to rise back the original values of the fields it is necessary to refer to a translation table which contains the record's structure. In such a table a triplicated dimension of [FieldOrdinal – FieldIndex - FieldValue] definition have to be compiled for each field of the record profile. For instance :

$$9.1) \quad \mathbf{R} = \{\text{City}=(Rome, Milan, Turin), \text{Age}=(20-25, 26-30, 31-35, 36-40), \text{Job}=(clerk, nurse, teacher)\}.$$

Expression is intentionally reminiscent of the one cited in point 4.1). The reference table would also be reported as the following T1 table :

Field Ordinal	Field Index	Field Value
1	1	Rome
1	2	Milan
1	3	Turin
2	1	20-25
2	2	25-30
2	3	<b>30-35</b>
2	4	35-40
3	1	clerk
3	2	nurse
3	3	teacher

We shall denote with  $\mathbf{S}$  the reference matrix structure related to T1 and with  $\mathbf{S}(i, j)$  the  $i$ th indexed field's value with index  $j$ ; thus in T1 it will be  $\mathbf{S}(2,3) = (30-35)$  (bolded line).

This formalism of  $\mathbf{S}$  it is very close to the real software modeling; programmatically, a bi-dimensional array  $\mathbf{S}(n, j; \mathbf{Max})$  can be handled with code of any programming language and a simple routine.

In a further step the original database table is expressed as table which lists a bi-dimensional array with the couple of significant columns: FRI and IDRec. The IDRec is clearly functional to the identification of the record in the original table, whereas the column of FRI will be used for the recursive RMM.

Pose  $\Gamma$  the bi-dimensional array containing the coupled values (FRI, IDRec). Each couple replace the record structure with the collection of fields in the table of origin. Table  $\Gamma$  will be called Factorial Table or FT because it consists of list of FRI which will be used for filtering, selecting, grouping and scanned throughout the matching algorithm.

Let  $\Lambda$  the FRI of a specific record  $R$  with  $n$  fields, the index  $\hat{j}(j)$  of the  $j$ th indexed field  $C_j$ ,  $j=1,..,n$  inside the record  $R$  we shall have :

$$9.2) \quad \hat{j}(j) = \log K_n(j) \frac{\Lambda}{\prod_{i=1..n, i \neq j} K_n(i)^{\hat{j}(i)}}$$

The computation for index  $\hat{j}(j)$  is recursive and all the others  $\hat{j}(i)$ , for  $i=1,..,n$ ,  $i \neq j$  are known. For a practical software implementation, it is also possible to use the index with multiple division operators instead of logarithmic functions which impact on CPU, it will be sufficient to calculate the number of times that  $\Lambda$  is dividable for  $K_n(j)$ . By extending the calculation of the indexes for all the fields of the record  $R$  it will be possible to achieve its entire content expressed with FRI

Let  $\Lambda$  the FRI and  $S$  the structure FT used for the record  $R$  where  $n$  fields were classified in an heuristic table, we shall have:

$$9.3) \quad R = \{(v_1), \dots, (v_n)\} \text{ with } v_i = S(i, \hat{j}(i)) \text{ for each } i=1,..,n.$$

## 10. SOFTWARE MODELING AND USAGE

For an easily applicable method it is essential to design and render an automated prototype with a software application. The paragraph 9.0 prospects an almost unlimited way of coding the logics of the RMM algorithm. In fact, there are several possible ways of combining dataset source, programming compilers, editing tools; besides all possible programmatic implementation an easily usable end-user interface have to be suitable to physician, nurses and so on. One privileged recommendation for programmers is specifically demanding efforts on the visual front-end for the user to be able to train and calibrate the FT for the indicator involved pending on the specific discipline.

On theory, to provide a basic modelling instruction a single functional modelling equation can be assumed thank to a fundamental algorithm, nonetheless for a generalised scheme and the largest possible applicability, would be strongly recommend an implementation that avoid hard-coded rules. Heuristic tables with classified fields values and relative matrices with relative weighted indexes will be handled outside the programmed calculation logics module; very likely it will reside in a simple editable file.

The RMM, cumulatively considered, can be shortened as follows :

$$10.1 \quad \Psi = (\underline{M}, \Omega, S, \Gamma, F)$$

where  $F$  represents the set of functions which account for the numerical computation, and several  $2n + 2$  arrays containing the  $\underline{M}_i$  and the  $\Omega_i$  structures are underneath.  $S$  is the structure which hosts the heuristic tables and  $\Gamma$  is the FT. Lastly,  $F$  it represents the functions library which will take care of the  $\Lambda$ ,  $\underline{K}$ ,  $\underline{P}$   $i$  and  $\hat{j}$  computation. A range  $\underline{K} < \delta$  will be defined for matching significance and  $\delta$  value will acquaint the meaning of a *cut-off* below which the ML will be considered significant.

A software should show a panel for the user to flag the variables included in the matching process; when considered outside the program logics the user will simply set to zero the field relevance (see 5.5). This panel should also allow users to launch matching inference on variously combined variables subsets.

One further possible feature of the model is to save a set of already ruled out FT so that ExRM trials can process different database sharing multi-centre survey and audit experiments. A pilot laboratory can stress-out empirical criteria on a reference dataset and persist a *template file* with the optimal combination of variables weights. Later this template can be distributed.

An important variant of the latter scheme is to save a *meta-file* after having ran the ExRM on a subset of control record. This is extremely useful when determining a calibration of a system by starting from either pre-determined or patronage tolerance within desirable limits of auditing indicators.

In a long-term prospective usage, persisted rules could be re-analysed to develop updatable system which can even auto-learn with a self-referential algorithm.

Proximity and similarity concepts translated in numeric FRI arrays can be filtered and grouped for hierarchical and sibling purposes (Social networks, Markov models). This would provide a mean for not only surmise on data but also clustering taxon and phenotypes in a ways that very much recalls UPGMA, BNN and Cohonen unsupervised algorithms [14] techniques.

A final interesting software implementation it would also reverse the calculation direction and design. Suppose we already have several subset of sufficiently “*closed*” records. By reverting the use of the algorithm, a software could easily extract meta-rules of ExRM. If we are strongly confident in some *set* of information which are considered as absolute paradigm of an ideally reference situation, the RMM can find the best heuristics FT and variable values classification for which already records fit a comparison standard (best fitting reasoning).

## 11. MODEL DISCUSSION

In order to be use on practise, the RMM model and its implementation have to be trained; the essential role of the  $\mathbf{n}$  matrix  $\mathbf{M}_i$  and the  $\mathbf{n}$  tri-dimensional arrays  $\mathbf{\Omega}_i$  (FT external files) determine the pertinence and the resolution capability of the knowledge system. This phase gives at the same time the knowledgebase and the specialisation of the model so that it can be properly applied to a the specific discipline.

The most relevant advantage of the factorised RMM model is the high flexibility and generalisation. The effectiveness is not dependent on the nature of data measurements or field’s casting. Simple trained matrices that adequately assign weights and discerning logics, can be easily understood and used by the physicians.

Especially worth for clinics is the possibility of changing and calibrating the variables panel to improve their affinity based on the observational medical experience. In the authors experience, physicians learn quite quickly how to translate conceptual knowledge in terms of relationships and dependencies of structured heuristics files.

RMM is intuitive and simple in that the training it consists on three basic steps which are quite natural and logical for a by investigator :

- 1) choice of variables and indicators to be focused,
- 2) compiling possible values for each variable,
- 3) *inter-field* enhancing factor choice for variables with higher consideration.

It can be noted how the efficacy of the model is strongly conditioned by the operator’s choices and its empirical experience. This is not a drawback because investigators can compare their expectation and make common assumptions on heuristic rules files. As in any other mathematical and statistics foundation, the RMM model only provides decisional support and suggestive numbers while the expert is supposed to play definitive interpretation to decide. Decision are also useful to progressively rectify the RMM’s rules.

The optimal usage of the RMM software is primarily concern with the meta-informational structure which have to be competent and strictly related to the scientific criteria. It is therefore critical the usability furnished to the physician with the software front-end. Authors have created software application for different ambits [13, 14, 15] and the RMM appeared to return meaningful results pending on the way it was trained rather than on the content of datasets. Crucial it was to compare RMM’s indication with ordinary statistics analysis conducted in similar studies [7, 8, 9, 10].

Because of its generalised flexibility the ExRM can be profitably used for a wide variety of application such as :

***Statistical Process Control  
Data-mining,  
Record-matching,***

***Medical Decision Aid systems  
Quality Assurance and Audit  
ICT security and cryptology***

## 12. Conclusion

The model of ExRM and its software RMM application have been formalised to provide an easy alternative to sophisticated statistical computation applied to either experimental, empirical and medical data. When data need to be investigated for non-parametric analysis, clustering reasoning or epidemiological stratification a simple record's matching algorithm based on Unique Factorisation Domain approach can offer a great speculative tool for data analysis. Virtually any type of dataset and experiments can be processed, and for practical software implementation, examples of source code concerning the modelling discussed in this work is freely distributed by the authors to anyone who wish to realise the software toolkit.

In the future, multiple practical application on diverse disciplines would be desirable to foster the model resolution and create a common mathematical standard for epidemiological meta-analysis on patient information. The auspice is that several other groups, involved in different scientific fields, could adopt the RMM and test its efficacy.

## Literature

1. **Baker, Alan**, *A Concise Introduction to the Theory of Numbers*, Cambridge University Press, Cambridge, UK, 1984)
2. **Cavallero A., Reina S., Schito G.C.** - Post Antibiotic Effect induced by Ofloxacin in both gram-positivi and gram-negative bacteria. "Chemoterapia" Jul 1987.
3. **FG.M. Artin**, *Algebra*, Prentice Hall (1991)
4. **D. S. Dummit, R. M. Foote**, *Abstract Algebra*, Wiley (1999).
5. **Hanai T, Honda H.** Application of knowledge information processing methods to biochemical engineering, biomedical and bioinformatics fields. *Adv Biochem Eng Biotechnol.* 2004;91:51-73. Review. PMID: 15453192 [PubMed - indexed for MEDLINE]
6. **Jiawei Han and Micheline Kamber** *Data Mining – Concepts and Techniques – Morgan Kauffman Publisher - 2001*
7. **Pollera C.F., Ameglio F., Reina S.** - Changes in serum iron levels following very high dose of cisplatin. *Cancer Chemotherapy and Pharmacology* 1987
8. **Reina S., Debbia E.A., Schito G.C.** - Ciprofloxacin Induced Modulation of cellular growth in activated, normal and lymphoid established Cell Lines. The antimicrobial agent resistances: orin treatment and control. *abs 70, 25 5 1991*, Principato di Monaco.
9. **Reina S., Debbia E., Schito G.C.** - Evaluation of the post antibiotic effect induced by various antibiotics against Staphylococci and Enterococci. *A.A.M.J.* 1993
10. **Reina S., Debbia E.** - Genetic recombination by spheroplast fusion in *Escherichia coli* K12. *Cytobios* by The Faculty Press. 1993, 76 91-95 .
11. **Reina S., E.A. Debbia, G.C. Schito.** In Vitro Cellular Growth Modulation by quinolone conditioned medium. 93<sup>rd</sup> General Meeting, Atlanta, Georgia, USA. Session 120. Paper nu. I28.
12. **Reina S., Boeri E., Lillo F., Cao Y., Varnier E.O.** Automation in AIDS research and diagnostic activity: a Local Area Network with Standard Query Language. 7<sup>th</sup> European Edition of Conference on Advanced Technology for Clinical Laboratory and Biotechnology. - ATB '91 Nov 26-11-1991 B11.
13. **Reina S., Miozza F.** - Knowledge Data Base System for Twins study. *ACTA GENET MED ET GEMELLOL.* Ed. Mendel Institute, Rome. 1994. 43:83-88
14. **Reina S., Reina V. , Giacomini M., Debbia E.** - Bio-fouling and micro-organisms identification on polluted materials: a novel Knowledge Data Base System architecture for a heuristic expert system engine. *Atti congresso Internazionale dei Biologi*, 22-25 settembre 1994. Vieste
15. **Reina S.** Il percent growth rate average (PGRA) migliora l'interpretazione dell'effetto post-antibiotico. 16mo Congresso AMCLI Nov 12-15 1987.
16. **Ruggiero C., Giacomini M., REINA. S., Gaglio S.** A qualitative process theory based model of the HIV-1 Virus-Cell interaction. *Proceedings of Medical Informatics Europe 93*, Israel. ISBN 965-294-091-7 pp. 147-150.
17. **Salvo A. Reina , Vito M. Reina and Eugenio A. Debbia** Records Matching model for data survey on applied and experimental microbiology (Submitted to *Annals of Microbiology*)
18. **Salvo A. Reina , Vito M. Reina , Mauro Costa and Alessandro Fasciani** - Novel applied software model for survey and trend analysis of endometriosis patients according to Evidence Based Medicine criteria. [Submitted to *Journal of Gynecology and Endocrinology*]
19. **Steven Roman**, *Coding and information theory*, Springer-Verlag, 1992
20. **Rotman J. J.**, *An introduction to the theory of groups - 4th ed*, Springer-Verlag, *Grad Texts in Math* 148, 1995