The ImageCLEF Medical Retrieval Task at ICPR 2010 — Information Fusion to Combine Visual and Textual Information

Henning Müller[1,2], Jayashree Kalpathy–Cramer[3]

[1]Geneva University Hospitals and University of Geneva, Switzerland
[2]University of Applied Sciences Western Switzerland, Sierre, Switzerland
[3]Oregon Health and Science University, Portland, OR, USA
http://www.imageclef.org/
henning.mueller@sim.hcuge.ch

Abstract. An increasing number of clinicians, researchers, educators and patients routinely search for medical information on the Internet as well as in image archives. However, image retrieval is far less understood and developed than text–based search. The ImageCLEF medical image retrieval task is an international benchmark that enables researchers to assess and compare techniques for medical image retrieval using standard test collections. Although text retrieval is mature and well researched, it is limited by the quality and availability of the annotations associated with the images. Advances in computer vision have led to methods for using the image itself as search entity. However, the success of purely content–based techniques has been limited and these systems have not had much clinical success. On the other hand a combination of text– and content–based retrieval can achieve improved retrieval performance if combined effectively. Combining visual and textual runs is not trivial based on experience in ImageCLEF. The goal of the fusion challenge at ICPR is to encourage participants to combine visual and textual results to improve search performance. Participants were provided textual and visual runs, as well as the results of the manual judgments from ImageCLEFmed 2008 as training data. The goal was to combine textual and visual runs from 2009. In this paper, we present the results from this ICPR contest.

1 Introduction

Image retrieval is a burgeoning area of research in medical informatics [1–3]. With the increasing use of digital imaging in all aspects of health care and medical research, there has been a substantial growth in the number of images being created every day in healthcare settings. An increasing number of clinicians, researchers, educators and patients routinely search for relevant medical information on the Internet as well as in image archives and PACS (Picture Archival and Communication Systems) [1, 3, 4]. Consequently, there is a critical need to manage the storage and retrieval of these image collections. However, image

retrieval is far less understood and developed than text–based searching. Text retrieval has a long history of evaluation campaigns in which different groups use a common test collection to compare the performance of their methods. The best known such campaign is the Text REtrevial Conference (TREC[1], [5]), which has been running continuously since 1992. There have been several offshoots from TREC, including the Cross–Language Evaluation Forum (CLEF[2]). CLEF operates on an annual cycle, and has produced numerous test collections since its inception in 2000 [6]. While CLEFs focus was originally on cross–language text retrieval it has grown to include multimedia retrieval tracks of several varieties. The largest of these, ImageCLEF[3] , started in 2003 as a response to the need for standardized image collections and a forum for evaluation. It has grown to become todays pre–eminent venue for image retrieval evaluation.

The coming sections will describe the ImageCLEF challenge itself and the details for the fusion task that was organized at ICPR (International Conference on Pattern Recognition). Then, the results and techniques of the participants will be analyzed in more detail and the main lessons learned from this context will be explained.

2 The Annual ImageCLEF Challenge

ImageCLEF is an international benchmark that includes several sub–tracks concerned with various aspects of image retrieval [7]; one of these tracks is the medical retrieval task run since 2004. This task within ImageCLEF enables researchers to assess and compare techniques for medical image retrieval using standard collections. ImageCLEFmed uses the same methodology as information retrieval challenges including TREC. Participants are given a set of topics that represent information needs. They submit an ordered list of runs that contain images that their system believe best meet the information need. Manual judgments using domain experts, typically clinicians, are used to create ground truth. The medical image retrieval tracks test collection began with a teaching database of 8,000 images in 2004. Since then, it has grown to a collection of over 74,000 images from the scientific literature, as well as a set of topics that are known to be well–suited for textual, visual or mixed retrieval methods. A major goal of ImageCLEF has been to foster development and growth of multimodal retrieval techniques: i.e., retrieval techniques that combine visual, textual, and other methods to improve retrieval performance.

Traditionally, image retrieval systems have been text–based, relying on the textual annotations or captions associated with images. Several commercial systems, such as Google Images[4] and Yahoo! images[5], employ this approach. Although text–based information retrieval methods are mature and well researched,

[1] http://trec.nist.gov/
[2] http://www.clef-campaign.org/
[3] http://www.imageclef.org/
[4] http://images.google.com/
[5] http://images.yahoo.com/

they are limited by the quality of the annotations applied to the images. Advances in techniques in computer vision have led to a second family of methods for image retrieval: content–based image retrieval (CBIR). In a CBIR system, the visual contents of the image itself are represented by visual features (colors, textures, shape) and compared to similar abstractions of all images in the database. Typically, such systems present the user with an ordered list of images that are visually most similar to the sample (or query) image. The text–based systems typically perform significantly better than purely visual systems at ImageCLEF.

Multimodal systems combine the textual information associated with the image with the actual image features in an effort to improve performance, especially early precision. However, our experience from the ImageCLEF challenge, especially of the last few years has been that these combinations of textual and visual systems can be quite fragile, with the mixed runs often performing worse than the corresponding textual run. We believe that advances in machine learning can be used more effectively to learn how best to incorporate the multimodal information to provide the user with search results that best meet their needs [8]. Thus, the goal of the fusion challenge at ICPR is to encourage participants to effectively combine visual and textual results to improve search performance. Participants were provided textual and visual runs that were submitted to the actual competition, as well as the results of the manual judgments from the ImageCLEFmed 2008 challenge as training data. The goal was to combine similar textual and visual runs from the 2009 challenge for testing. In this paper, we present the preliminary results from this ICPR contest.

3 The ImageCLEF Fusion Challenge

In both 2008 and 2009, the Radiological Society of North America (RSNA[6]) made a subset of its journals image collections available for use by participants in ImageCLEF. The 2009 database contains 74,902 images, the largest collection yet [9]. The organizers created a set of 25 search topics based on a user study conducted at Oregon Health & Science University (OHSU) in 2009 [4]. These topics consisted of 10 visual, 10 mixed and 5 semantically oriented topics, as categorized by the organizers based on past experience and nature of the query. During 2008 and 2009, a panel of clinicians, using a web–based interface, created relevance judgments. The manually judged results were used to evaluate the submitted runs using the trec_eval[7] software package. This package provides commonly used information retrieval measures including mean average precision (MAP), recall as well as precision at various levels for all topics.

For the ICPR fusion contest, the goal was to combine the best visual and textual runs that had been submitted previously to improve performance over the purely visual and purely textual runs. After participants registered they were provided access to the training data in early November 2009. The training set consisted of the four best textual and visual runs from different groups in 2008.

[6] http://www.rsna.org/
[7] http://trec.nist.gov/trec_eval/

Only one of these groups participated in the fusion challenge, so there was no advantage for any group. These runs were anonymized to remove information about the group. We also provided the qrel, the file that contained the output for the manual judgments as well as the results obtained by the training runs using the trec_eval package. Participants could create fusion runs using combinations of the provided training runs and evaluate the performance using the trec_eval along with the abovementioned qrel file as well as the results of the evaluation measures for the runs.

We released the test runs two weeks later. Again these consisted of the four best textual and four best visual runs, this time from 2009. The ground truth in the form of qrel was not provided at this time. The judgments were released in early January so that the participants could evaluate their runs in time for submission to ICPR 2010. To summarize, the timeline for this contest was as follows:

- 16.11.2009 Release of training data
- 30.11.2009 Release of test data
- 04.01.2010 Submission of results
- 10.01.2010 Release of ground truth data
- 15.01.2010 Conference paper submission

4 Fusion Techniques used by the Participants

There was quite a variety of techniques relying on either the similarity scores of the supplied runs or the ranks. Early fusion was hardly possible as only the outcome of the system was supplied and no further information, limiting the variety of the approaches.

OHSU used a simple scheme based principally on the number of times that a particular image occurs in the results sets as the main criterion. Two runs use only textual information (fusion2, fusion4) and two runs combine both visual and textual techniques (fusion1, fusion3). Then as a second criterion either the sum of the ranks was used (fusion1, fusion2) or the sum of the scores (fusion3, fusion4).

The *MedGIFT* (Medical projects around the GNU Image Finding Tool) group employed two principal approaches for the fusion described in more detail in [10]. Methods are based on ranks and on the scores. Whereas ranks can be used directly, the scores were normalized to be in the range 0..1 to be better comparable among the submissions. In terms of combination rules a max combination was used where of all systems the maximum was taken (combMax), a sum rule summing up normalized scores or ranks (combSum) and the last rule includes the frequency of the documents into this (CombMnz).

The results of the best system (*SIFT*) in the context are described in [11]. This group uses a probabilistic fusion, where weights are calculated from training data (ProbFuse). The training takes into account that documents retrieved later are generally less relevant and these are subsequently weighted with a learned decrease of the weight (SegFuse). All these techniques can create border effects as

the results are grouped in blocks. This can be removed with SlideFuse. SlideFuse most often had the best results.

The *PRISMA* group developed two methods called rankMixer and rankBooster. Both take into account the frequency of an image in the results lists to be combined and its scores. These are used to calculate a function for calculating the similarity score for a particular image.

Finally, the *ISDM* group developed an approach based on a generative statistical model. It uses an attentive reader model, meaning that early documents are weighted high and then attention decreases, in their case with a logarithmic model. The importance of single runs is in a second approach estimated based on population–based incremental learning.

5 Results of the Participants

Table 1 contains the performance of the training runs that were provided. As can be seen, the textual runs perform significantly better than the visual runs for all measures. This has to be taken into account when combining the runs.

Table 1. Results of the training runs.

Run	Recall	MAP	P5	P10
Text1	0.63	0.29	0.49	0.46
Text2	0.65	0.28	0.51	0.47
Text3	0.54	0.27	0.51	0.47
Text4	0.61	0.28	0.44	0.41
Visual1	0.06	0.028	0.15	0.13
Visual2	0.24	0.035	0.17	0.17
Visual3	0.17	0.042	0.22	0.17

This performance gap was similarly true for the test runs (Table 2). Overall, the performance was better for the textual runs in 2009 whereas it was worse for the visual runs as can be seen when comparing the two tables.

Participants were successful in creating fusion runs that were better than the original text and visual runs, as well being substantially better than the official mixed runs that had been submitted to ImageCLEFmed 2009. None of the officially submitted fusion runs was better than the best text run in the competition.

We received 49 runs from five groups as part of the fusion task. Of the 35 mixed runs that were submitted, 18 had higher MAP compared to the best textual training run and interestingly, 25 had higher MAP compared to the best official mixed run in 2009 as seen in Figure 1. This shows the potential performance gains through fusing varying techniques and it shows how little focus most ImageCLEF participants put into this..

Table 2. Results of the test runs.

Run	Recall	MAP	P5	P10
Text1	0.73	0.35	0.58	0.56
Text2	0.66	0.35	0.65	0.62
Text3	0.77	0.43	0.70	0.66
Text4	0.80	0.38	0.65	0.62
Visual1	0.12	0.01	0.09	0.08
Visual2	0.12	0.01	0.08	0.07
Visual3	0.11	0.01	0.09	0.07
Visual4	0.11	0.01	0.09	0.08

Figure 2 shows the precisions of the best original runs and the best fusion runs. There is a slight improvement in early precision with the best fusion runs both textual and mixed. However, the fusion runs created using only visual runs performed quite poorly, which is not surprising as the basic results were all very low. Although there was little difference between the best fusion mixed and textual runs for the MAP, the runs with highest early precision used the visual runs in combination with the textual runs. This underlines the importance of visual information, even with a very poor performance, for early precision. This also shows that the information contained in visual and textual retrieval runs is very complementary.

In Table 3 the results when fusing only the textual runs are shown. The best runs of each performance measure are marked in bold. Best results are obtained with a probabilistic model that learned the importance of specific parts of the results. The best four results are all very close. MAP and early precision are both very well correlated among the runs and the best run regarding MAP also had best early precision. BPref (Binary preference) shows whether a technique has many un–judged images ranked highly and in this case it correlates very closely with MAP, which is not surprising as all runs are based on the exact same runs or basic technqiues.

Table 4 shows the visual fusion results of the participants. Only a single group submitted three runs. Results could be increased over the original results but they remained low as the based results were not performing well at all. Other participants also combined visual runs only without submitting them to the contest but results were very similar to the results presented here.

Table 5 displays all submitted mixed runs. The early precision and the MAP of these runs are clearly superior to all the text runs shown in Table 2. We can also see that the best runs in terms of MAP are not best in terms of early precision, so to understand these a more detailed analysis of the techniques needs to be performed. All among the early results have a very similar score. The first six runs only have an absolute difference in terms of MAP of 1%. When compared to the fusion results using only text it can be seen that MAP is slightly lower but early precision is significantly lower with a much higher margin.

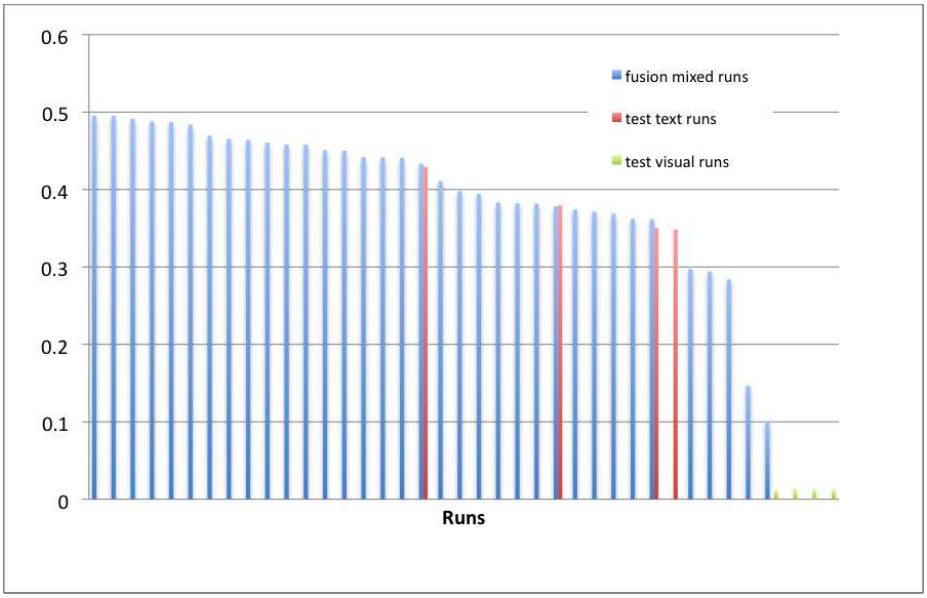

Fig. 1. MAP of all fusion runs and test runs.

Combinations of only the textual runs delivered similar results to the mixed runs with the best technique (SIFT group) obtaining 0.487, so slightly lower than the combination of the mixed runs. Other groups similarly had slightly better results using the mixed combinations compared to only comparing the text runs. For early precision this was similar but with a stronger difference, obtaining 0.72 compared to 0.76 for the best mixed combination run, with most other groups having a slightly lower early precision for the text only runs.

6 Conclusions

The first fusion challenge to combine visual and textual runs from medical image retrieval was organized for ICPR 2010. The goal of this context was to encourage participants to explore machine learning and other advanced techniques to effectively combine runs from the ImageCLEFmed challenge given a set of training runs and their performance metrics. Five groups submitted a total of 49 runs, many of which demonstrated the effectiveness of a multimodal approach to image retrieval. It was encouraging to note that about half of the submitted runs performed better than all the test runs. On the other hand, a few of the mixed runs that we submitted performed poorly, possibly due to the really poor performance of the visual test runs. The best runs obtained a MAP of 0.495 compared to the best run in the ImageCLEF of 0.43 and the best combined run in ImageCLEF 2009 of even 0.41. Such gains of over 20% show the potential of well combining visual and textual cues for medical image retrieval. The focus of Im-

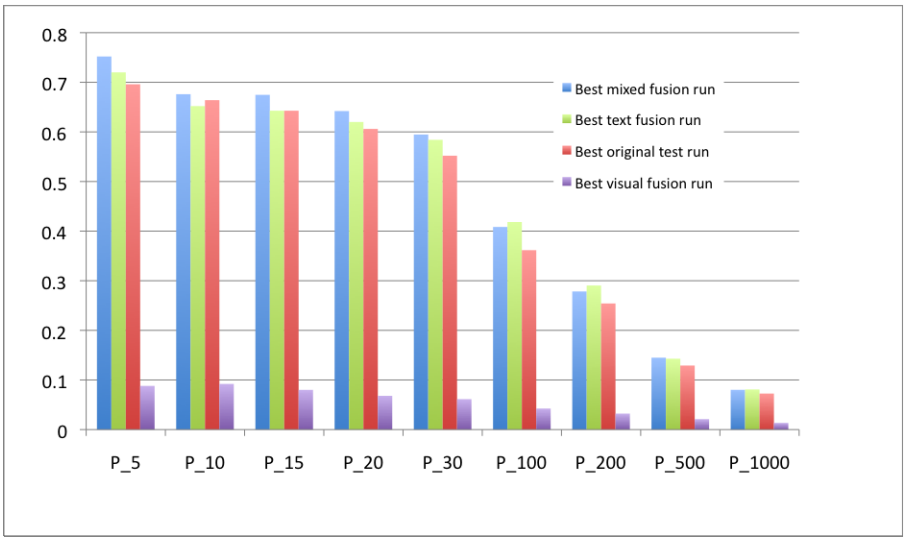

Fig. 2. Early precision (P_X meaning precision after X documents are retrieved) of original text runs and fusion runs.

ageCLEF should be on fostering such developments In the past, particularly the combination of media has been of limited effectiveness in ImageCLEF as most research groups work on either visual or textual retrieval but not the two. The small participation of only five research groups on the other hand also showed that there might be even more potential if successful techniques for fusion are consistently applied and tested.

7 Acknowledgements

We would like to acknowledge the support of the National Library of Medicine grant 1K99LM009889 and of the BeMeVIS project of the HES–SO.

References

1. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medicine–clinical benefits and future directions. IJMI **73**(1) (February 2004) 1–23
2. Tagare, H.D., Jaffe, C., Duncan, J.: Medical image databases: A content–based retrieval approach. JAMIA **4**(3) (1997) 184–198
3. Hersh, W., Müller, H., Jensen, J., Yang, J., Gorman, P., Ruch, P.: Advancing biomedical image retrieval: Development and analysis of a test collection. JAMIA **13**(5) (September/October 2006) 488–496
4. Radhouani, S., Kalpathy-Cramer, J., Bedrick, S., Hersh, W.: Medical image retrieval, a user study. Technical report, Medical Inforamtics and Outcome Research, OHSU, Portland, OR, USA (June 2009)

Table 3. Performance metrics for fusion of text runs.

Group	Runid	type	map	bpref	P5	P10	P30
SIFT, Ireland	txtOnlySlideFuse	Textual	**0.487**	**0.499**	**0.72**	0.652	0.584
PRISMA, Chile	testt234v	Textual	0.480	0.498	0.704	**0.672**	0.5973
PRISMA, Chile	testt1234v	Textual	0.474	0.487	0.712	0.648	0.596
PRISMA, Chile	testt123v	Textual	0.473	0.489	0.712	0.664	0.584
SIFT, Ireland	txtOnlySegFuse	Textual	0.466	0.472	0.696	0.668	0.577
PRISMA, Chile	testt124v	Textual	0.464	0.474	0.688	0.64	**0.604**
SIFT, Ireland	txtOnlyProbFuse	Textual	0.447	0.454	0.704	0.652	0.556
PRISMA, Chile	testt134v	Textual	0.43	0.444	0.712	0.656	0.563
OHSU, USA	fusion1	Textual	0.300	0.337	0.448	0.376	0.381
OHSU, USA	fusion4	Textual	0.270	0.347	0.28	0.332	0.361
OHSU, USA	fusion3	Textual	0.175	0.235	0.328	0.32	0.24

Table 4. Performance metrics for fusion of visual runs.

Group	Runid	type	map	bpref	P5	P10	P30
SIFT, Ireland	visOnlySegFuse	Visual	**0.0179**	0.0353	0.088	**0.092**	0.0613
SIFT, Ireland	visOnlySlideFuse	Visual	0.0175	**0.0354**	**0.104**	0.088	**0.064**
SIFT, Ireland	visOnlyProbFuse	Visual	0.0154	0.0338	0.088	0.08	0.056

5. Harman, D.: Overview of the first Text REtrieval Conference (TREC–1). In: Proceedings of the first Text REtrieval Conference (TREC–1), Washington DC, USA (1992) 1–20
6. Savoy, J.: Report on CLEF–2001 experiments. In: Report on the CLEF Conference 2001 (Cross Language Evaluation Forum), Darmstadt, Germany, Springer LNCS 2406 (2002) 27–43
7. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: CLEF 2007 Proceedings. Volume 5152 of Lecture Notes in Computer Science (LNCS)., Budapest, Hungary, Springer (2008) 473–491
8. Müller, H., Kalpathy-Cramer, J.: Analyzing the content out of context — features and gaps in medical image retrieval. International Journal on Healthcare Information Systems and Informatics **4**(1) (2009) 88–98
9. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Said, R., Bakke, B., Kahn Jr., C.E., Hersh, W.: Overview of the CLEF 2009 medical image retrieval track. In: Working Notes of CLEF 2009, Corfu, Greece (September 2009)
10. Zhou, X., Depeursinge, A., Müller, H.: Information fusion for combining visual and textual image retrieval. In: Pattern Recognition, International Conference on, Los Alamitos, CA, USA, IEEE Computer Society (2010)
11. Zhang, L., Toolan, F., Lillis, D., Collier, R., Dunnion, J.: Probabilistic data fusion for image retrieval. In: Pattern Recognition, International Conference on, Los Alamitos, CA, USA, IEEE Computer Society (2010)

Table 5. Performance metrics for fusion mixed runs.

Group	Runid	type	map	bpref	P5	P10	P30
SIFT, Ireland	txtimgSlideFuse	Mixed	**0.495**	0.494	0.712	0.66	0.588
SIFT, Ireland	txtimgSlideFuse	Mixed	**0.495**	0.494	0.712	0.66	0.588
PRISMA, Chile	gt841t234v3	Mixed	0.491	**0.497**	**0.76**	**0.696**	**0.611**
medGIFT, CH	combSUMlogRank	Mixed	0.488	0.490	0.712	0.672	0.592
medGIFT, CH	combMNZlogRank	Mixed	0.487	0.489	0.712	0.672	0.592
medGIFT, CH	combSUMByFreqlogRank	Mixed	0.484	0.489	0.712	0.672	0.592
SIFT, Ireland	txtimgSegFuse	Mixed	0.469	0.459	0.696	0.672	0.585
PRISMA, Chile	testt1234v234	Mixed	0.466	0.461	0.752	0.676	0.595
PRISMA, Chile	testt1234v134	Mixed	0.464	0.458	0.744	0.692	0.589
medGIFT, CH	GESUM3MAXLinearRank	Mixed	0.461	0.465	0.720	0.656	0.556
OHSU, USA	fusion2	Mixed	0.458	0.478	0.672	0.628	0.575
medGIFT, CH	SUM3MAXByFreqLinearRank	Mixed	0.458	0.462	0.720	0.656	0.556
PRISMA, Chile	testt1234v123	Mixed	0.451	0.449	0.744	0.688	0.577
PRISMA, Chile	testt1234v124	Mixed	0.450	0.449	0.712	0.656	0.576
medGIFT, CH	combMNZScoreNorm	Mixed	0.442	0.442	0.720	0.656	0.579
medGIFT, CH	combSUMFreqScoreNorm	Mixed	0.442	0.446	0.688	0.692	0.568
medGIFT, CH	combSUMScoreNorm	Mixed	0.441	0.442	0.720	0.656	0.579
SIFT, Ireland	txtimgProbFuse	Mixed	0.434	0.419	0.696	0.652	0.563
PRISMA, Chile	testt1234v1234	Mixed	0.411	0.403	0.720	0.648	0.551
PRISMA, Chile	testt134v234	Mixed	0.398	0.399	0.664	0.664	0.528
PRISMA, Chile	testt134v134	Mixed	0.394	0.395	0.688	0.66	0.528
ISDM, Spain	gen2	Mixed	0.383	0.385	0.688	0.668	0.54
ISDM, Spain	gen5	Mixed	0.382	0.384	0.696	0.652	0.536
ISDM, Spain	gen1	Mixed	0.382	0.385	0.688	0.664	0.54
ISDM, Spain	gen4	Mixed	0.379	0.382	0.704	0.66	0.527
PRISMA, Chile	testt234v124	Mixed	0.374	0.373	0.616	0.604	0.523
ISDM, Spain	gen3	Mixed	0.372	0.375	0.688	0.64	0.521
PRISMA, Chile	testt134v123	Mixed	0.369	0.374	0.648	0.6	0.509
PRISMA, Chile	testt134v124	Mixed	0.362	0.374	0.616	0.596	0.508
PRISMA, Chile	testt124v123	Mixed	0.362	0.357	0.608	0.596	0.517
PRISMA, Chile	testt234v1234	Mixed	0.298	0.304	0.520	0.508	0.431
PRISMA, Chile	testt134v1234	Mixed	0.294	0.304	0.520	0.5 00	0.437
PRISMA, Chile	testt124v1234	Mixed	0.284	0.291	0.496	0.504	0.420
ISDM, Spain	wsum1	Mixed	0.147	0.200	0.528	0.456	0.293
ISDM, Spain	wmnz1	Mixed	0.100	0.142	0.352	0.292	0.216