Ground Truth Generation in Medical Imaging

A Crowdsourcing–based Iterative Approach

Antonio Foncubierta-Rodríguez[*]
University of Applied Sciences Western
Switzerland
TechnoArk 3
3960 Sierre, Switzerland
antonio.foncubierta@hevs.ch

Henning Müller
University of Applied Sciences Western
Switzerland
TechnoArk 3
3960 Sierre, Switzerland
henning.mueller@hevs.ch

ABSTRACT

As in many other scientific domains where computer–based tools need to be evaluated, also medical imaging often requires the expensive generation of manual ground truth. For some specific tasks medical doctors can be required to guarantee high quality and valid results, whereas other tasks such as the image modality classification described in this text can in sufficiently high quality be performed with simple domain experts.

Crowdsourcing has received much attention in many domains recently as volunteers perform so–called human intelligence tasks for often small amounts of money, allowing to reduce the cost of creating manually annotated data sets and ground truth in evaluation tasks. On the other hand there has often been a discussion on the quality when using unknown experts. Controlling task quality has remained one of the main challenges in crowdsourcing approaches as potentially the persons performing the tasks may not be interested in results quality but rather their payment.

On the other hand several crowdsourcing platforms such as Crowdflower that we used allow creating interfaces and sharing them with only a limited number of known persons. The text describes the interfaces developed and the quality obtained through manual annotation of several domain experts and one medical doctor. Particularly the feedback loop of semi–automatic tools is explained. The results of an initial crowdsourcing round classifying medical images into a set of image categories were manually controlled by domain experts and then used to train an automatic system that visually classified these images. The automatic classification results were then used to manually confirm or refuse the automatic classes, reducing the time for the initial tasks.

Crowdsourcing platforms allow creating a large variety of interfaces for judgements. Whether used among known experts or paying for unknown persons, they allow increasing the speed of ground truth creation and limit the amount of money to be paid.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Crowdsourcing, information classification and retrieval, ground truth

1. INTRODUCTION

The amount of medical images daily produced grows rapidly both in clinical and scientific environments [12]. Classifying these images into categories is the entry point to efficient retrieval as it allows limiting the search space. In clinical domains, images are often used only once for diagnosis purposes of a single patient. Other usage of the images is possible if they could easily be accessed, for example for training young physicians or as clinical decision support [5, 6]. In scientific environments, retrieval of relevant documents plays an important role for researchers. Medical image retrieval and medical image classification [12, 1] have been used as a way of improving access to visual medical information. Medical retrieval tasks have attracted the interest of many researchers. The image retrieval benchmarking event ImageCLEF has included a medical task since 2004 and a medical image modality classification task [10] since 2010. Imaging modalities initially included mainly clinical image types such as computed tomography (CT) or x–rays but also dermatology images and various types of graphs. In 2012, a hierarchy of over 30 modality types was created to well group images for the retrieval.

Evaluation campaigns or benchmarking events require a consistent ground truth for accurate and reliable evaluation of the participating systems or methods. The medical image modality classification in ImageCLEF currently contains a dataset of more than 300'000 images from the open access biomedical literature. Obtaining a full ground truth for this enormous dataset is an expensive and time consuming task. Evaluation of automatic methods is possible but requires a consistent, representative training set to produce a reliable classification of the data. Once some training data are available an automatic approach can be used, and manual in-

[*] Corresponding author.

CrowdMM'12, October 29, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1589-0/12/10 ...$15.00.

tervention would only require validation with a *right/wrong* selection, which can mean a faster and simpler task.

In this paper we present a set of experiments aimed at evaluating the use of crowdsourcing for ground truth generation in modality classification. Domain experts and physicians are compared and then also external judges are used with a strict quality control. The experiments show the potential of crowdsourcing for several tasks to obtain quick and inexpensive results. It also shows that a two step approach can have advantages.

2. METHODS

In this section the crowdsourcing platform chosen for the experiments is presented. The details of the tasks are explained, with a focus on the user groups and phases of the experiments.

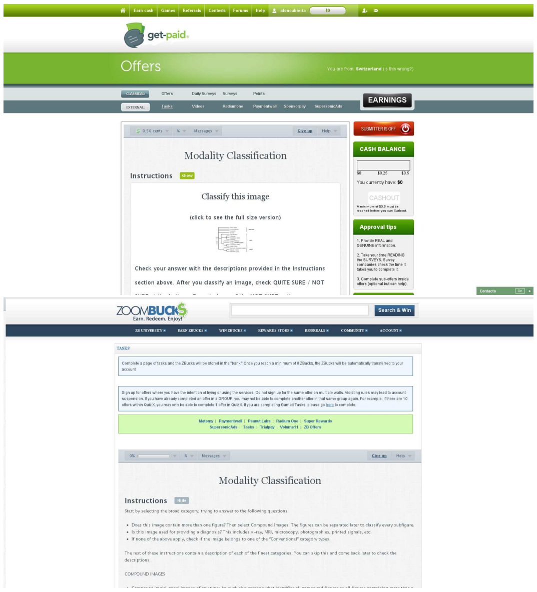
2.1 Finding the right crowdsourcing platform

Several platforms are available for outsourcing tasks to an undefined group of people often referred as *the crowd* [8]. Amazon Mechanical Turk[1], GetPaid[2], ZoomBucks[3] and others offer the possibility of managing and creating jobs consisting on small tasks that users complete for a small amount of money per task. For instance, Amazon Mechanical Turk was previously used by the Information Retrieval (IR) community either for annotation [2, 14] or relevance evaluation [3]. In the medical domain, Amazon Mechanical Turk has recently been used by Nguyen et al. [13] for Computer Aided Diagnosis.

The choice of the crowdsourcing platform may limit the extent of the experiment in terms of the target workforce, the task design, or even geographical limitations for creating and advertising the job. Crowdflower[4] functions as a platform hub, allowing the job designer to use the crowdflower specific markup language to build the task GUI and redistributing/advertising the jobs in several other crowdsourcing platforms. This allows, for example, to obtain judgements from several geographic regions and several user profiles without having to reimplement the tasks. Another interesting feature is that is the provision of a free internal interface that can be used for small to mid–scale evaluation of the interface or for comparing crowd–based results to results provided by a known set of people. Examples of our interfaces are shown in Figure 1.

2.2 Details of the task to be performed

Manual annotation in free text is prone to spelling errors, typos and definition misunderstandings that would result in the addition of an enormous amount of work in the quality control phase. Therefore a hierarchy of 34 categories [11] was used in order to provide a reference for classifying a part of the 300'000 images. Figure 2 shows some samples of these categories. The hierarchical nature of the categories allows for analyzing up to which point there is agreement across annotators. It is possible to identify points where further specification is required or which categories can be merged. The categories considered for our annotation are shown in Table 1, where the 34 categories are the leaf nodes in bold.



You are from: **Switzerland** (is this wrong?)

Modality Classification

Instructions

Classify this image

(click to see the full size version)

Check your answer with the descriptions provided in the Instructions section above. After you classify an image, check QUITE SURE / NOT

You currently have: **$0**

A minimum of $0.5 must be reached before you can Cashout.

1. Provide REAL and GENUINE information.
2. Take your time READING the SURVEYS. Survey companies check the time it takes you to complete it.
3. Complete sub-offers inside offers (optional but can help).

Complete a page of tasks and the ZBucks will be stored in the "bank." Once you reach a minimum of 8 ZBucks, the ZBucks will be automatically transferred to your account!

Sign up for offers where you have the intention of trying or using the services. Do not sign up for the same offer on multiple walls. Violating rules may lead to account suspension. If you have already completed an offer in a GROUP, you may not be able to complete another offer in that same group again. For example, if there are 10 offers within Quiz X, you may only be able to complete 1 offer in Quiz X. If you are completing Gambit Tasks, please go here to complete.

Modality Classification

Instructions

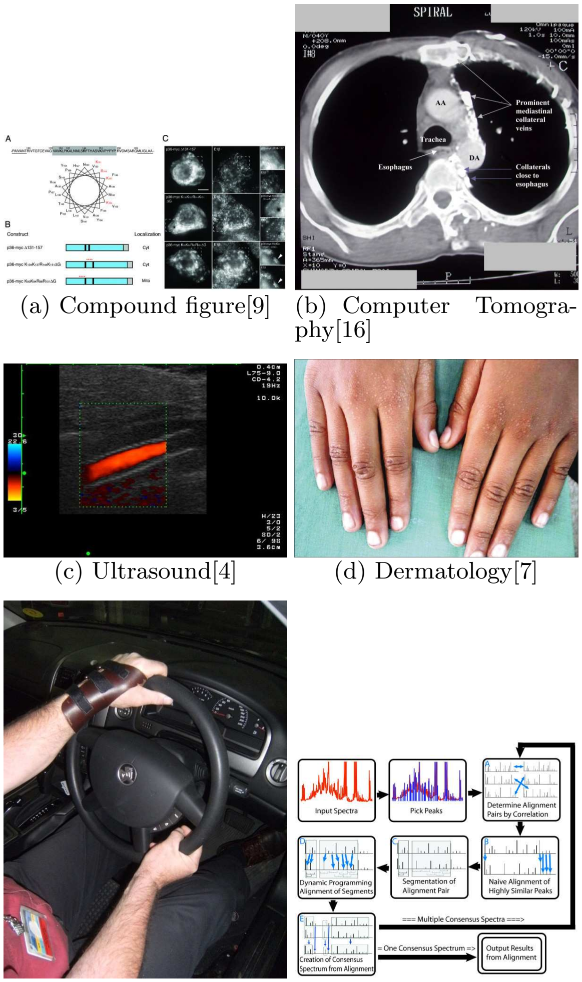
Start by selecting the broad category, trying to answer to the following questions:

- Does this image contain more than one figure? Then select Compound Images. The figures can be separated later to classify every subfigure.
- Is this image used for providing a diagnosis? This includes x–ray, MRI, microscopy, photographies, printed signals, etc.
- If none of the above apply, check if the image belongs to one of the "Conventional" category types.

The rest of these instructions contain a description of each of the finest categories. You can skip this and come back later to check the descriptions.

COMPOUND IMAGES

Figure 1: Examples for the crowdflower–generated interfaces used in our tests.


A
B
C
Construct
Localization
SPIRAL
AA
Trachea
Esophagus
DA
Prominent mediastinal collateral veins
Collaterals close to esophagus
Input Spectra
Pick Peaks
Determine Alignment Pairs by Correlation
Dynamic Programming Alignment of Segments
Segmentation of Alignment Pair
Naive Alignment of Highly Similar Peaks
Creation of Consensus Spectrum from Alignment
Multiple Consensus Spectra
One Consensus Spectrum
Output Results from Alignment

(a) Compound figure[9]
(b) Computer Tomography[16]
(c) Ultrasound[4]
(d) Dermatology[7]
(e) Non–clinical photo[17]
(f) System overview[15]

Figure 2: Sample images from several categories.

[1] http://www.mturk.com/
[2] http://www.get-paid.com/
[3] http://www.zoombucks.com/
[4] http://www.crowdflower.com/

Table 1: Hierarchical image modalities.

Compound / multi–panel images of any type			
Diagnostic Imaging	Radiology images	**Ultrasound/echo**	
		Magnetic Resonance Imaging, MRI	
		Computerized Tomography, CT	
		2D Radiography, film or digital	
		Angiography, radiography of vessels with a contrast agent	
		Positron Emission Tomography, PET	
		Single Photon Emission Computed Tomography, SPECT	
		Combined modalities in one image such PET/CT, PET/MRI, dual energy CT, fMRI	
		Infrared	
	Visible light photography, gross level	Gross photography of organs, tissue	**Skin**
			Other organs
		Endoscopy pictures	
	Printed Signals, waves	**EEG**	
		ECG / EKG	
		EMG	
	Microscopic images	**Light microscopy**	
		Electron microscope	**Transmission microscope**
		Fluorescence images	
		Microscopy, interference	**Phase contrast**
		Dark field	
	Reconstructed, rendered images	**3D reconstructions or 3D views**	
		2D reconstructions	
Conventional biomedical illustrations	Graphs	**Tables, forms**	
		Program listing	
		Statistical figures, graphs, pie charts, histograms, other charts ...	
		Screenshots	
		Flow charts	
		System overviews or overviews of components including links and graphics for the parts	
		Gene sequence	
		Chromatography, Gel	
		Chemical structure	
		Symbol	
		Mathematics, formulae	
	Non clinical photos		
	Hand–drawn sketches		

The ground–tuthing task was divided into several steps that were executed in an iterative way.

Initial training set generation.

The first task performed was aimed at obtaining an initial training set of 1000 images by using the internal crowdsourcing interface by 18 known users. All users were familiar with medical imaging and the modality hierarchy, with experience in the medical imaging domain varying from 1 year to a decade.

Automated classification and verification.

Once a small training set was manually labelled, the complete set of 300'000 images was automatically classified using a visual words approach and the training set as reference. Then, a second crowdsourcing task was set up for simply validating or refusing the automatically assigned class. This allows for a faster annotation of correctly classified images and reduces the amount of images that need to be reclassified.

Results trustability.

In order to evaluate the expected accuracy for different user groups, three additional experiments were executed with three user groups:

- A 45–year–old medical doctor (MD) was asked to classify a set of 3415 images. A random selection of 1661 images out of these were used as gold standard for measuring trustability of the two remaining user groups.
- A known set of experts familiar with the modality hierarchy, all researchers in the medical imaging domain with experience from one to more than ten years, were asked to reclassify part of the 3415 images classified by the MD. The images were presented in groups of two. One of the judgments was used for measuring agreement with the medical doctor, informing the user of the possible errors. In case of disagreement, the user had the chance of contesting the gold standard providing additional information on the ambiguity of the classes and misjudgements of users.
- The same experiment was performed advertising the task in several crowdsourcing platforms. The 1661 gold standard images were divided into 2 groups: 814 public gold images and 847 hidden gold images, that do not alert the user of the errors. This was done to reduce the influence of gold images on the remaining judgments, while keeping a large amount of gold units to be used as a trustability threshold.

3. RESULTS

In this section the results of the various experiments are presented.

3.1 User self–assessment

For each judgement, the user was required to answer how sure he/she was of the choice. Even a simple method like this can be valuable for discarding single judgements from people that did not feel confident about their choice. Instead

	High confidence jugdements
Medical doctor	100%
Known experts	95.04%
Crowd (all)	85.56%

Table 2: User self–confidence value.

	Agreement
Broad Category	88.76%
Diagnostic subcategory	97.40%
Microscopy	89.06%
Radiology	90.91%
Reconstructions	100%
Visible light photography	79.41%
Conventional subcategory	76.95%

Table 3: Agreement between a MD and known experts for various categories.

of discarding all the information from not trustable sources, the self–confidence value allows keeping the data users are confident with. Table 2 shows the percentage of judgements with high self–confidence.

3.2 Crowdsourcing with known experts compared to an MD

Table 3 shows the agreements for the hierarchy tree. Disagreements existed particularly within conventional imaging where the subcategories can contain an important ambiguity.

Known experts and an MD also produce annotations at a different speed. The MD classified 3415 images at a rate of 85 judgements per hour. The group of known experts classified fewer images at a slower rate: 66 judgements per hour. Figure 3 shows the amount of images classified per contributor in the known group. The experiment was open to the complete contributor group during the same period of time.

3.3 Open crowdsourcing compared with an MD

Since there was a quality control threshold during the open crowdsourcing experiment several of the contributors were rejected after only few judgements. The amount of trusted judgements achieved in one week was 10463 whereas the amount of judgements from non–trusted contributors was 15706. This difference also affects the distribution of the other results. For fair comparison with the closed crowdsourcing experiment, only trusted judgements are considered.

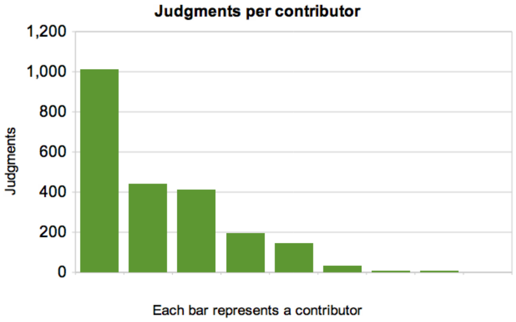


Figure 3: Judgements in the group of known persons.

	Agreement rate
Broad Category	85.53%
Diagnostic subcategory	85.15%
Microscopy	70.89%
Radiology	64.01%
Reconstructions	0%
Visible light photography	58.89%
Conventional subcategory	75.91%

Table 4: Agreement for various modalities when comparing the MD to the open crowdsourcing contributors.

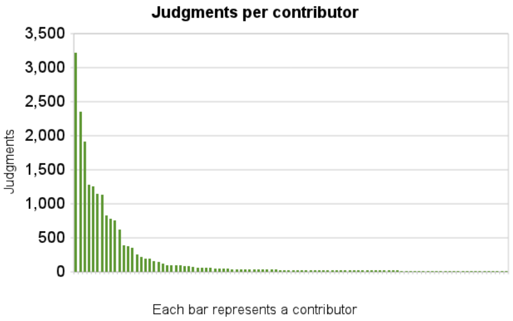


Figure 4: Judgements in the open experiment.

Similarly to results from section 3.2, Table 4 shows the agreement in the hierarchy tree. Agreement between the judgements and the physician are all lower than the agreement of the known experts with the MD. Particularly in diagnostic imaging this difference is high and these categories are most important. Figure 4 shows the top 100 contributors. In terms of speed, the external contributors were on average much slower than the medical doctor or the known group, providing a rate of 25 judgements per hour. However, the enormous amount of contributors, more than 2470, increases the amount of judgements that can be received, even if not all of them are fully trusted.

3.4 Confirming automatically classified results

The results from a first round of images classified with a known expert group, were used for training an automated system for classifying the remaining images. Out of the first 1000 images verified, the agreement among annotators was 100%. Annotators were assessed by means of manually-generated ground truth randomly presented to users. In terms of speed, verification takes much less effort to the user, being able to answer almost twice as fast as in the full annotation experiment. In our first tests some mistakes in the machine classification led to an accuracy of only 24%. The usefulness is clearly higher, the better the automatic classification is as this reduces the amount of images that need to be reclassified.

4. DISCUSSION AND CONCLUSIONS

This articles describes the use of crowdsourcing for generating ground truth for ImageCLEF 2012. The modality hierarchy of 34 classes into which images from the biomedical literature needed to be classified required specific knowledge but a detailed description was made available for the judges to explain the classes. Crowdflower was used as crowdsourcing platform as it is possible to be used from Europe and also because interfaces can be developed for internal tasks without payment.

Results show that there is an important difference in the trustability and behaviour expected from a closed group and *the crowd*. Due the size of the groups, the crowd provides much faster results, often at a cost of trustability in some categories. Specifically, the external group was able to distinguish the broad categories and could accurately distinguish the various diagnostic subcategories with an important difference between judges. However, the finer–grained classification was not very accurate, even though the judgements considered are only those from contributors with more than 70% overall accuracy compared to a given gold standard. The amount of contributors as well as the required accuracy per contributor can be a key to obtaining better annotations. In an article published during the execution of this experiment, Nguyen et al. [13] obtain an accuracy comparable to Computer Aided Detection (CAD) with a reduced number of contributors (150 in the first trial, and 102 in the second trial) with 95% or above approval rating in the Amazon Mechanical Turk platform. However, using the approval rating based on previous and not necessarily related tasks can be misleading in terms of trustability. A higher threshold based on the actual task, by using a gold standard as explained in Section 2.2, might be a compromise between the two approaches.

The use of a strategy in several steps, creating first a training set and then mainly having experts confirm or reject the system response can help to reduce the time and money necessary for manual judgements. So far we only executed two steps but a similar approach can be repeated, retraining an automatic system successively and thus increasing the automatic classification accuracy.

Crowdsourcing has shown to be an extremely useful tool to obtain ground truth of good quality using manual intelligence tasks. Quality control is essential to obtain good results as the quality of the judges for the task can vary strongly. For extremely large data sets, e.g. hundreds of thousands of images, crowdsourcing is the best possible method in terms of speed and accuracy if none or insufficient training data is available for using a computer–based approach. It seems important to make sure that tasks are quick to perform and simple as quality challenges occurred particularly in the more difficult categories. It also needs to be noted that even domain experts can not agree on the exact modality of all images, so a certain subjectivity will remain in any case.

5. ACKNOWLEDGMENTS

We would like to thank the European Union for their funding in the context of Khresmoi (grant number 257528), Chorus+, and the Swiss National Science foundation (grant number 205321–130046).

6. REFERENCES

[1] C. Akgül, D. Rubin, S. Napel, C. Beaulieu, H. Greenspan, and B. Acar. Content–based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging*, 24(2):208–222, 2011.

[2] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *SIGIR 2009 Workshop on The Future of IR Evaluation*, 2009.

[3] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, December 2008.

[4] C. Arning and U. Grzyska. Color doppler imaging of cervicocephalic fibromuscular dysplasiaâĂć. *Color Doppler imaging of cervicocephalic fibromuscular dysplasia*, 2, 2004.

[5] B. Caputo, H. Müller, T. S. Mahmood, J. Kalpathy-Cramer, F. Wang, and J. Duncan. Editorial of miccai workshop proceedings on medical content–based retrieval for clinical decision support. In *Proceedings on MICCAI Workshop on Medical Content–based Retrieval for Clinical Decision Support*, volume 5853 of *Lecture Notes in Computer Science (LNCS)*. Springer, 2009.

[6] A. Depeursinge, H. Greenspan, T. Syeda Mahmood, and H. Müller. Overview of the second workshop on medical content-based retrieval for clinical decision support. In H. Greenspan, H. Müller, and T. Syeda Mahmood, editors, *Medical Content-based Retrieval for Clinical Decision Support*, MCBR–CDS 2011. Lecture Notes in Computer Sciences (LNCS), Sept. 2011.

[7] S. Gupta, S. D. Shenoi, and V. Mehta. Acquired Lymphangioma Of Vulva Secondary To Radiotherapy For Carcinoma Cervix. *Indian Journal of Dermatology*, 53(4):221–222, 2008.

[8] J. Howe. Crowdsourcing. `http://www.crowdsourcing.com/cs`, 2012. [Online; accessed 29-Jun-2012].

[9] Y. T. Hwang, A. W. Mccartney, S. K. Gidda, and R. T. Mullen. Localization of the Carnation Italian ringspot virus replication protein p36 to the mitochondrial outer membrane is mediated by an internal targeting signal and the TOM complex. *BMC Cell Biology*, 9:54+, 2008.

[10] J. Kalpathy-Cramer, H. Müller, S. Bedrick, I. Eggel, A. G. Seco de Herrera, and T. Tsikrika. The CLEF 2011 medical image retrieval and classification tasks. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, September 2011.

[11] H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, and S. Antani. Creating a classification of image types in the medical literature for visual categorization. In *SPIE medical imaging*, 2012.

[12] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content–based image retrieval systems in medicine–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.

[13] T. B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, J. E. Burns, and R. M.

Summers. Distributed human intelligence for colonic polyp classification in computer–aided detection for CT colonography. *Radiology*, 262(3):824–828, March 2012.

[14] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, New York, NY, USA, 2010. ACM.

[15] J. Staab, T. O'Connell, and S. Gomez. Enhancing metabolomic data analysis with Progressive Consensus Alignment of NMR Spectra (PCANS). *BMC Bioinformatics*, 11(1):123+, 2010.

[16] H. Tang and J. H. K. Ng. Googling for a diagnosis — use of Google as a diagnostic aid: Internet based study. *British Medical Journal*, 333:1283–1284, 2006.

[17] J. Thiele, R. Nimmo, W. Rowell, S. Quinn, and G. Jones. A randomized single blind crossover trial comparing leather and commercial wrist splints for treating chronic wrist pain in adults. *BMC Musculoskeletal Disorders*, 10:129+, 2009.