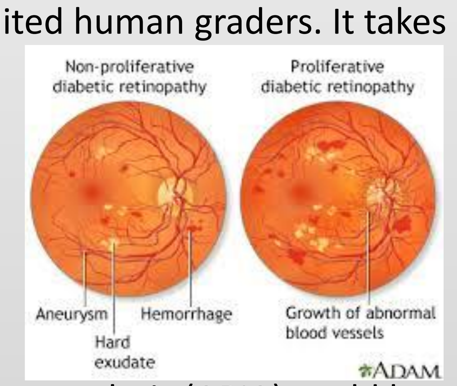
ARIA or Manual?- A statistical analysis into which method of grading retinal images for diabetic retinopathy is the most effective

Introduction

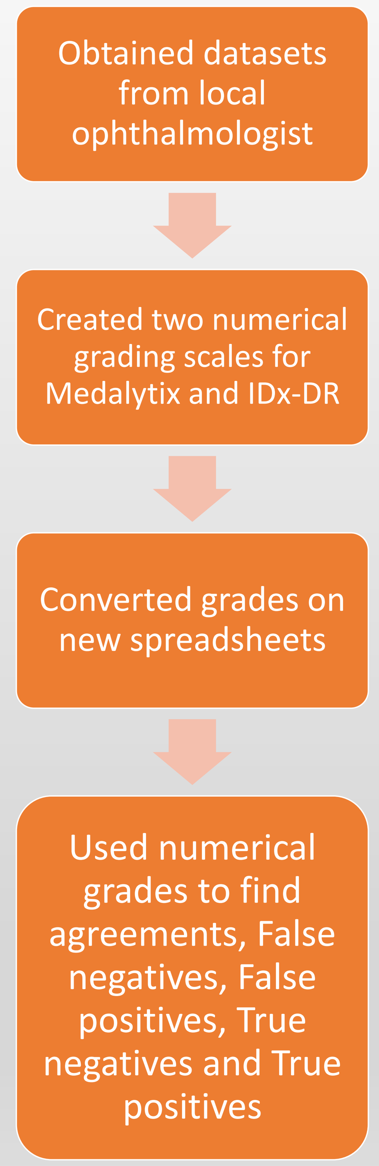
Diabetes is a global epidemic affecting 422 million people worldwide. A common complication of diabetes is diabetic retinopathy. This is a condition of the small vessels in the retina and presents in different stages. In 2011, 93 million people worldwide suffered from diabetic retinopathy. In Ireland alone, 1 person a week goes blind due to this condition (diabetes.ie).

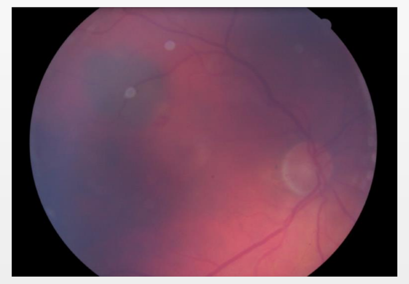
Diabetic retinopathy is detected through screening. The most common method of obtaining retinal images is fundus photography. Normally, images obtained from screening are analysed manually by trained, accredited human graders. It takes a human grader 1.5 times longer to decide an image is normal than to spot disease (Tang,2014). This is because the markers of disease can be very obvious e.g haemorrhages, hard exudates, cotton wool spots, microaneurysms and growth of new blood vessels. Automated retinal image analysis (ARIA) could be used as a workload reduction tool in the grading of retinal images. An automated system that could safely reduce the number of ‘no disease cases’ would make the grading process more efficient and cost effective



Our aim is to compare the IDx DR and Medalytix automated systems grades to the gold standard of trained accredited human graders.

Data

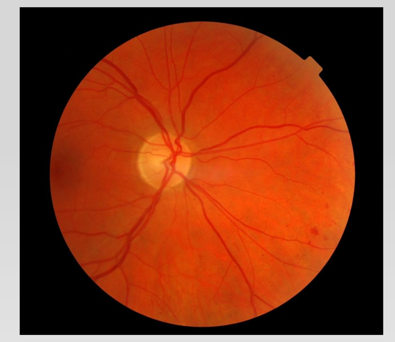




Medalytix case
Inadequate quality

IDx-DR case:
68719476786
No retinopathy

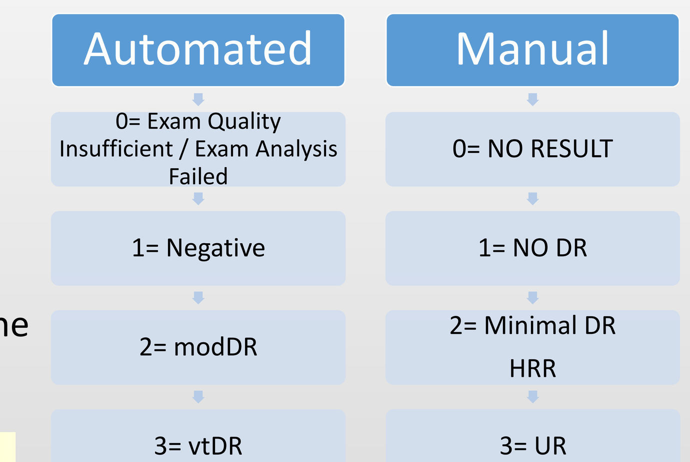
IDx-DR case:
68719476804
Retinopathy present due to red lesions and scarring near right fovea



Changed case
IDx-DR case:
68719476797
IDx grade- Mod DR
Manual grade- UR
IDx was in fact right

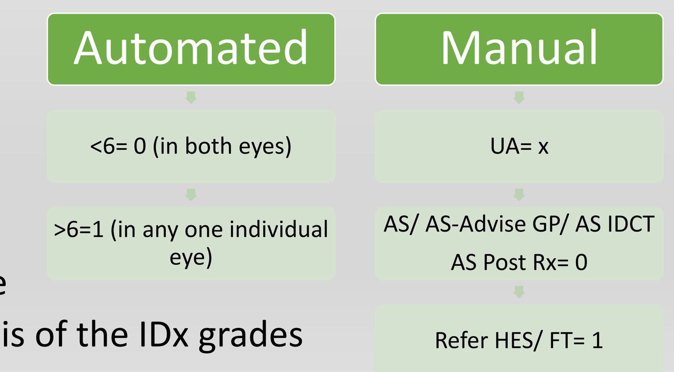
Materials and methods

We obtained two datasets from a local ophthalmologist in the form of Excel spreadsheets. The first dataset consisted of 95 cases, all of which were Type 2 Diabetic, Irish patients. Their retinal images had been graded by trained accredited Irish human graders and also by the IDx-DR automated system. Similarly the second dataset contained 535 cases, all of which were Irish patients. Their retinal images were graded by the same Irish human graders and the Medalytix automated system. We began with our IDx-DR dataset and created a new Excel spreadsheet containing the two grades; from the manual graders and the automated system. We then changed these grades to numerical grades based on the scale seen here. We used Excel operations to find and sort the agreements, false positives, true positives, false negatives, true positives and true negatives. From these figures we used the following formulae to find the IDx-DR statistical values.



$\text{Accuracy}=\frac{TP+TN}{total\ pop.}$ $\text{NPV}=\frac{TN}{TN+FN}$ $\text{PPV}=\frac{TP}{TP+FP}$ $\text{Specificity}=\frac{TN}{TN+FP}$ $\text{Sensitivity}=\frac{TP}{TP+FN}$

Similarly, we created a new spreadsheet from the original Medalytix dataset. We also converted the automated and manual grades to numbers using another grading scale (shown here). We then used the same operations to find the systems agreements, true positives and negatives, false positives and negatives. We subsequently found the Medalytix automated systems sensitivity, specificity, accuracy, PPV and NPV using the equations as above. After analyzing the IDx spreadsheet, an ophthalmologist reviewed the results and changed some cases, so we changed the statistical values accordingly. We also did another analysis of the IDx grades using a grading scale more similar to the one we used for the Medalytix data for additional results.



Literature cited

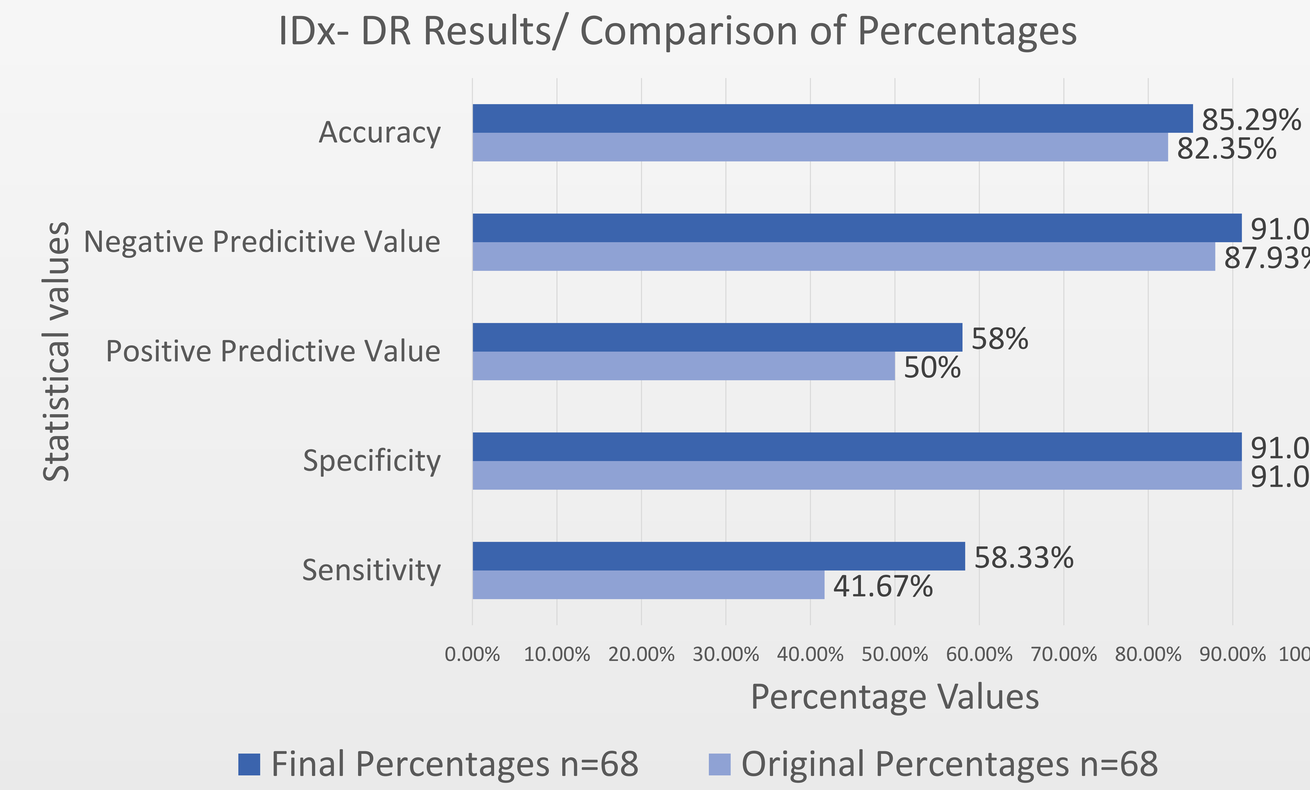
Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." *Jama* 316.22 (2016): 2402. Web.

"Home." *Diabetes Ireland*. N.p., n.d. Web. 04 Jan. 2017.

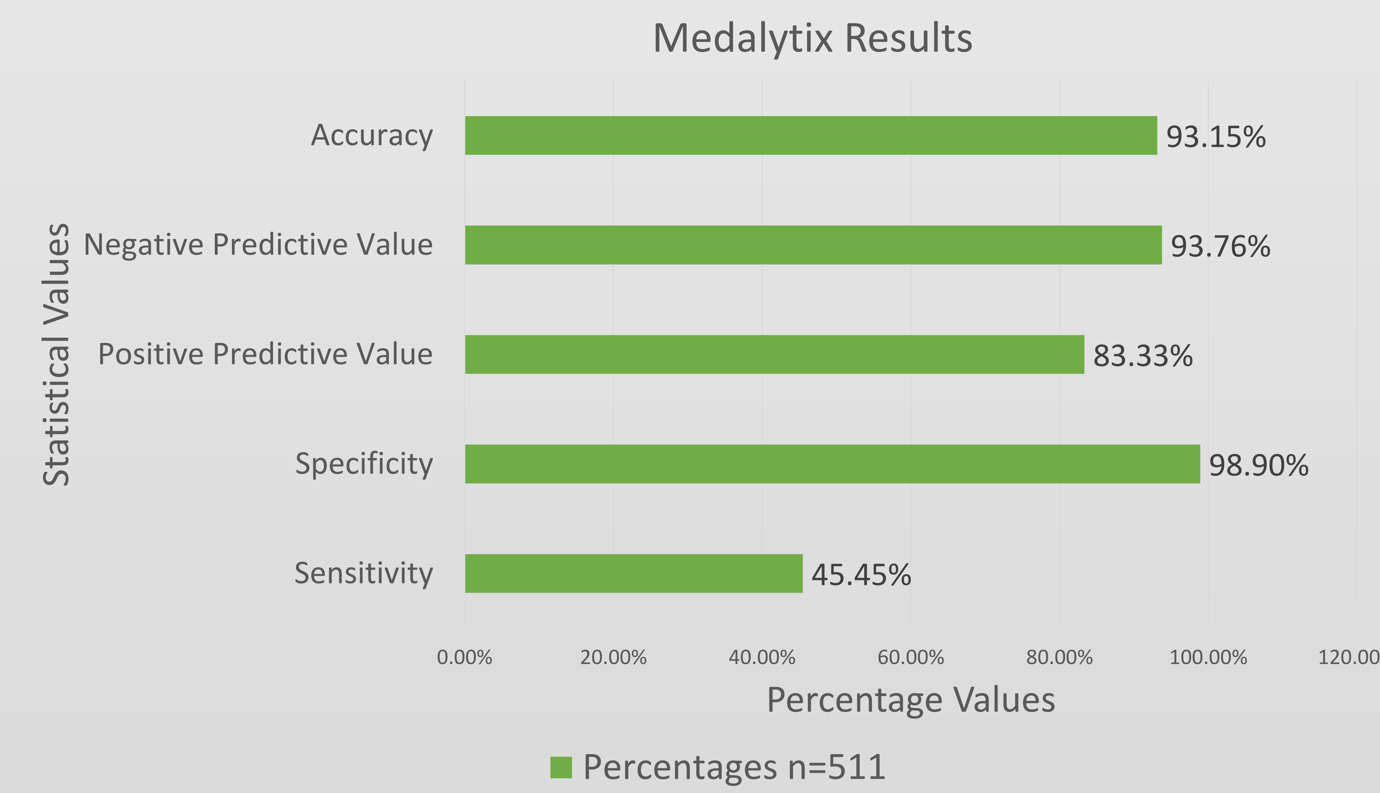
Tang, Hongying Lilian, Jonathan Goh, Tunde Peto, Bingo Wing-Kuen Ling, Lutfiah Ismail Al Turk, Yin Hu, Su Wang, and George Michael Saleh. "Correction: The Reading of Components of Diabetic Retinopathy: An Evolutionary Approach for Filtering Normal Digital Fundus Imaging in Screening and Population Based Studies." *PLoS ONE* 9.1 (2014): n. pag. Web.

Results

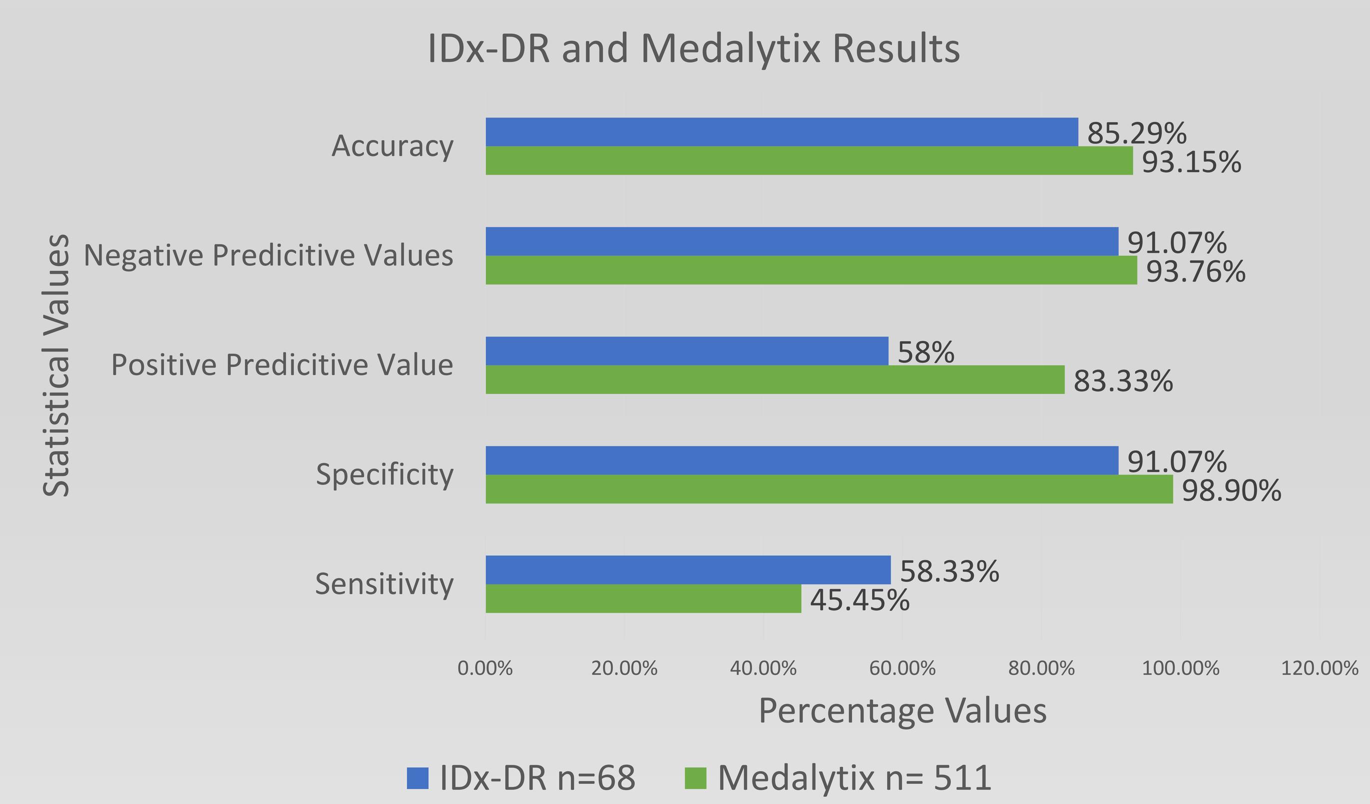
We had two sets of results for the IDx-DR data due to the review changing the results of certain cases



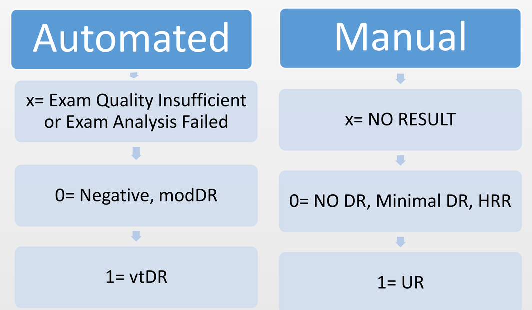
Similarly for the Medalytix data, we found all of the same values. This data was not reviewed therefore there was only one set of results.



Here is a comparison of both sets of results

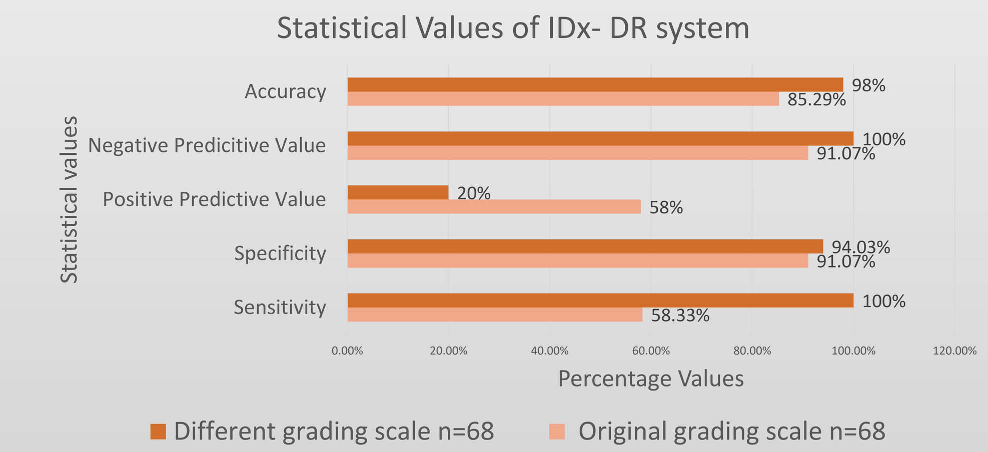


Additional Results



We went on to reanalyze the IDx results using a similar grading scale to that of Medalytix.

This meant that less cases were regarded as positive.



All of the values increased other than the PPV which is due to the lack of positive cases. The 100% sensitivity and 94% specificity mean that this system would be adequate for completing the task of eliminating negative disease cases from the manual grading pool.

Conclusions

To conclude our results show that both IDx-DR and Medalytix can be used as an effective method of removing “no disease” images from the manual grading queue. The negative predictive values of both systems are over 93% and the specificities of both are over 91% demonstrating that both systems can identify “true negative” cases well.

However, with relatively low sensitivities (both below 59%) it is questionable if an automated system would be entirely effective and clinically safe in the grading process of urgent referral cases. The original sensitivities and specificities of both systems don’t meet the requirements of the WHO and British Diabetic Association. Alternatively, in the grading process of minimal diabetic retinopathy cases a lower sensitivity and specificity may be acceptable.

Medalytix did appear to be the most effective at grading of the two systems from our comparative statistical graph as all of the statistical values found from Medalytix, other than the sensitivity, were higher than those of IDx-DR. This could be due to the fact that the protocol Medalytix is modeled on (NSC) is similar to the protocol the manual graders are trained on (ENSPDR) in comparison to the protocol IDx follows (ICDR). Also, the grading scale we used for the Medalytix spreadsheet may have caused the results to be higher in general as having only two grades made it easier for results to agree. When we used this grading scale on the IDx dataset, we found that this helped results agree also and gave very high sensitivities and specificities.

Overall an automated system would be a cost effective, time saving and safe alternative to identifying negative disease cases manually. A solution to the low sensitivity of the systems would be human review of all positive cases. The deep learning feature on automated systems could help improve their performance into the future (Gulshan) Our suggestion is for automated retinal image analysis to be used in synergy with manual graders The increasing burden of diabetic screening will only worsen as the epidemic of diabetics grows in Ireland and worldwide.