Harnessing citizen science through mobile phone technology to screen for immunohistochemical biomarkers in bladder cancer

Authors: [1]Peter Smittenaar*, [2]Alexandra K Walker*, [2]Shaun McGill*, [3,4]Christiana Kartsonaki, [1]Rupesh J Robinson-Vyas, [1]Janette P McQuillan, [1]Sarah Christie, [1]Leslie Harris, [1]Jonathan Lawson, [2]Elizabeth Henderson, [5]Will Howat, [6]Andrew Hanby, [7]Gareth J Thomas, [8]Selina Bhattarai, [9]Lisa Browning, [2]Anne E Kiltie

Address:
[1] Cancer Research UK, 407 St John Street, EC1V 4AD, London, UK
[2] CRUK/MRC Oxford Institute for Radiation Oncology, University of Oxford, OX3 7DQ, UK
[3] Department of Population Health, University of Oxford, OX3 7LF, UK
[4] MRC Population Health Research Unit, University of Oxford, OX3 LF, UK
[5] Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 ORE, UK
[6] Leeds Institute of Cancer and Pathology (LICAP), St James's University Hospital Beckett Street Leeds, LS9 7TF
[7] Cancer Sciences Unit, University of Southampton Faculty of Medicine, Tremona Road, Southampton SO16 6YD, UK
[8] Leeds Teaching Hospitals NHS Trust, St James's Hospital, Leeds LS7 9TF, UK
[9] Department of Cellular Pathology, and the NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, OX3 9DU, UK.

*These authors contributed equally to this work.

Running title: Mobile crowdsourcing of biomarker analysis

Correspondence:
Prof Anne E Kiltie
CRUK/MRC Oxford Institute for Radiation Oncology
Department of Oncology
University of Oxford
Old Road Campus Research Building
Off Roosevelt Drive
OXFORD OX3 7DQ

Tel. 01865 617352

anne.kiltie@oncology.ox.ac.uk

Word count: 5113
Number of figures: 5
Number of tables: 1

Abstract

Background

Immunohistochemistry (IHC) is often used in personalisation of cancer treatments. Analysis of large data sets to uncover predictive biomarkers by specialists can be enormously time-consuming. Here we investigated crowdsourcing as a means of reliably analysing immunostained cancer samples to discover biomarkers predictive of cancer survival.

Methods

We crowdsourced the analysis of bladder cancer TMA core samples through the smartphone app 'Reverse the Odds'. Scores from members of the public were pooled and compared to a gold standard set scored by appropriate specialists. We also used crowdsourced scores to assess associations with disease-specific survival.

Results

Data were collected over 721 days, with 4,744,339 classifications performed. The average time per classification was approximately 15 seconds, with approximately 20,000 hours total non-gaming time contributed. The correlation between crowdsourced and expert H-scores (staining intensity x proportion) varied from 0.65 to 0.92 across the markers tested, with six of 10 correlation coefficients at least 0.80. At least two markers (MRE11 and CK20) were significantly associated with survival in patients with bladder cancer, and a further three markers showed results warranting expert follow-up.

Conclusion

Crowdsourcing through a smartphone app has the potential to accurately screen IHC data and greatly increase the speed of biomarker discovery.

Key words (3-12 words): Citizen Science, crowdsourcing, bladder cancer, immunohistochemistry, tissue microarrays.

Introduction

Personalised medicine involves tailoring treatment to reflect a patient's individual tumour characteristics. For this to be used routinely, we need to find biomarkers robustly associated with cancer prognosis or predictive of outcome following therapy.

Immunohistochemistry (IHC) is widely used for biomarker identification. To automate the staining and analysis process, IHC is often combined with tissue microarray (TMA) technology. TMAs position hundreds of small-diameter tissue samples in a physical array, which can be stained and scanned as one unit. This has made it possible to generate a large volume of IHC data from many patients relatively quickly. The analysis of IHC stained tissue is largely performed by the naked eye. Such 'manual' scoring of IHC is time consuming and requires several trained researchers or histopathologists to reach a consensus score. Automated analysis software is quickly gaining ground especially for the most common cancers and stain types. Though there is little doubt such algorithms will eventually improve beyond human capability, currently automated IHC scoring algorithms are not applicable to all samples, and manual intervention is required for challenging or ambiguous cases[1,2,3]. A major barrier to more accurate algorithms is the availability of large labelled data sets to train new generations of supervised algorithms. Crowdsourcing is one commonly used approach to generate such labels, essentially facilitating, rather than competing with, algorithms.

Recently, a number of projects have reported success in using crowdsourcing for the analysis of large data sets within a range of scientific disciplines, including biochemistry and biomedical research[4,5,6,7,8,9,10]. 'Cell Slider' is one such study which aimed to address the rate-limiting step of IHC manual scoring (https://www.cellslider.net/). Here, untrained members of the public were able to accurately score IHC data, with participants achieving similar results to trained pathologists in cancer cell identification, oestrogen receptor (ER) status assignment and associations of ER status with clinical outcome in breast cancer.

However, participants demonstrated a bias in terms of overestimating the number of cancer cells in an image, thus compromising the accuracy of IHC scoring[8]. 'Trailblazer' was developed following Cell Slider to further test the ability of the public to score IHC data and to identify methodological improvements that could increase the accuracy of publicly generated scores. When given more comprehensive tutorials, the public was highly accurate in their detection of cancer and IHC scoring[10].

Issues with crowdsourcing include the drop-off rate and inactivity in user participation. Recent crowdsourcing projects have attempted to increase the audience base to increase participants' activity by integrating scientific tasks into games. Examples of such ventures include Foldit[5], Phylo[6], EteRNA[7] and Fraxinus[9]. Such crowdsourcing games have led to *bona fide* scientific discoveries and generated improvements in existing computational algorithms[11].

We identified a number of proteins worthy of assessment for potential associations with clinical outcome in lung and bladder cancer and conducted IHC on TMAs containing tissue from patient tumour samples. Due to the use of radiotherapy as a treatment modality in MIBC, a number of proteins involved in the repair of DNA damage were assessed (RAD50, MRE11, p53, p21). CK20 and CK5/6 were included in this study as these immunostains have previously been used to distinguish basal and luminal MIBC subtypes[12]. MRE11, p53, p21, TIP60 and Ki67 IHC have previously been investigated in MIBC and found to be potentially associated with survival and/or cancer progression[13,14,15,16,17,18,19,20]. However, there is currently not enough evidence to conclude whether these proteins are valid biomarkers for DSS in MIBC. We decided to employ crowdsourcing scoring to analyse the large amount of data generated. A key objective was to retain the accuracy of manual scoring while being less time-consuming for the experts.

Scoring IHC data in its most basic form is a task involving pattern recognition and determination of colour gradients. We therefore hypothesised that, given a short tutorial,

members of the public as a group would be able to accurately assess the IHC staining of cancerous tissue and that crowdsourcing could increase the speed of scoring large sets of IHC data. Unlike previous web-based crowdsourcing efforts in IHC scoring, we input our IHC data into a mobile gaming app available to members of the public. We first assessed the accuracy of the crowdsourced data. Then, when the crowdsourced scores were found to be accurate, we used these scores to look for associations between protein staining and clinical outcome.

Materials and Methods

User recruitment

We crowdsourced the analysis of bladder cancer TMA core samples through the smartphone application (app) Reverse the Odds (RTO), distributed through Google Play and the iTunes store. Users in the app classified TMA samples as described below, and every so often were offered a separate minigame based on Reversi. TMA classifications were incentivised through powerups that could be used in the minigame. As such, during TMA classification there were no distractions, however the minigame provided some variety to an otherwise highly repetitive task. The majority of time was spent performing TMA classifications, though exact figures are unavailable. We did not store personal information about the users nor information regarding which user provided each classification. The data reported here were collected between 9th October 2014 and 28th September 2016.

Tissue microarray samples

Ethical approval was obtained from London Bromley NRES (study 13/LO/0540), Leeds (East) Local Ethical Committee (studies 02/060 and 04/Q1206/62) and North West – Haydock Research Ethics Committee (study 14/NW/1033). Patients whose samples were collected from 2002 onwards gave informed consent for use of their pre-treatment biopsies.

All bladder cancer tissue cores were collected from four cohorts of patients. The first three were treated with radical radiotherapy at the Leeds Cancer Centre, UK (1995-2000, 2002-2005 and 2006-2009), and the remaining cohort was treated with radical cystectomy at the Leeds Teaching Hospitals NHS Trust, UK (1995-2005). The 1995-2005 cohorts have been previously described in Choudhury *et al* (2010)[21], and the 2006-2009 radiotherapy patients (n=47) were treated as per the 2002-2005 radiotherapy cohort.

Haematoxylin and eosin (H&E)-stained sections from formalin-fixed paraffin embedded bladder tumour samples, taken at pre-treatment transurethral tumour resection, were

reviewed by a consultant uropathologist (SB) and areas of invasive transitional cell carcinoma were outlined. Using a Beecher tissue microarrayer, 1356 0.6 µm cores were taken from up to five muscle-invasive areas per sample, and made into seven TMAs of up to 21 x 18 samples, including barrier samples of placenta, liver or mouse liver.

Immunohistochemical staining

Lung cancer samples were stained for scoring in Reverse the Odds using anti-CD8 and AntiPDL1 antibodies and were entered into the game but not analysed further due to small number of responses and a shift in focus towards bladder cancer.

For bladder cancer, 11 different IHC stains were tested, using a BOND autostainer or manual methods. Details of the staining are given in Supplementary methods, with specific antibody conditions listed in Table S1. Slides were scanned using an Aperio ScanScope CS2 digital slide scanner at x400 magnification and viewed using Aperio Image Scope viewing software. TMAs were then segmented using Aperio TMALab software. For use in the RTO app, the colours of the images were transformed from DAB and haematoxylin stained to inverted colours, to make scoring of the samples more appealing to the general public (Figure 1). Cores were split into 36 segments to allow the user to comfortably inspect individual cells. The running order is shown in Table S2 and the results for the bladder cancer samples presented.

Task design

Users were presented with a brief tutorial explaining how to spot cancerous tissue/cells, how to assess the proportion of cancer cells that were stained, and how to assess the intensity of staining (Figure S1A-C). They were then presented with a segment and asked to score it (Figure 1 and Figure S1D and E). We only asked for the proportion of staining if cancer was indicated in the first question, and we only asked for intensity of staining if proportion was indicated as >0%. For the ‘proportion of cancer cells stained’ question we used 5 numerical

ranges but used three different sets of ranges depending on the specific stain (Table 1): Category 1: 0, 1-25, 25-50, 50-75, 75-95, 95-100%; category 2: 0, 1-10, 10-25, 25-50, 50-75, 75-100%; category 3: 0, 1-10, 10-25, 25-65, 65-95, 95-100%. Users could access the instructions at any point during this process.

Gold standard scoring

Experts (AK and SMcG, with AW where consensus could not be achieved between two scorers) assessed the presence of cancer and proportion and intensity of staining in DAB-stained whole cores from TMAs to assess the performance of the crowdsourcing method. Gold standard scores were provided for whole TMA cores for approximately 10 to 15% of the data (Table 1).

Statistical analysis

Aggregation of scores

We aggregated individual scores to arrive at a single proportion and intensity score for each core. Each core was scored between 81 and 3676 times across all of its segments (mean: 405, SD: 455, median: 124, IQR: [97, 1030]). Rather than aggregate responses first by segment and then by core, we took the mean across all proportion and intensity scores respectively for a core, ignoring any responses where the participant indicated there was no cancer. When a response indicated “no cancer”, this might be because they were shown a piece of the TMA without cancer cells. In that case we were correct to exclude their response from the aggregate. If the user did erroneously report “no cancer” for a sample with cancer, their response was incorrect and we were also correct to exclude it. As the dataset in some cases contained multiple cores per patient, we combined cores for each patient by taking the mean for proportion and intensity.

Linear correction of scores

Taking the (weighted) arithmetic mean across users and then segments, as done here, is an appropriate way of averaging out noisy scores, but only if errors are symmetrically distributed around the true mean. However, because the scores are bounded in this experiment, we would expect to consistently overestimate scores that are close to zero, and underestimate scores that are close to the maximum. For example, if the true proportion of cancer cells stained by the marker is 0, then 'noisy' individual users can only *overestimate* the score. As such, there are no *underestimates*—because a user cannot provide an answer that is smaller than 0—and the average user score is biased in a positive direction, i.e. it can only be >=0.

To correct for this bias in proportion and intensity scores respectively, we applied a linear correction with clipping at the minimum and maximum values (e.g. any scores corrected by the linear model that ended up below 0 were set to 0). For the cores that had no expert scores, we could use all expert-scored cores to determine the intercept and slope between cores scored by both citizen scientists and experts. To correct the cores also scored by experts, we could not use the data that needed correction in the calculation of the very correction. This would lead to overly accurate crowdsourced scores. We therefore applied 10-fold cross-validation, using *sklearn.cross_validation.cross_val_predict,* described in Pedregosa *et al* (2011)[22] with a linear regression estimator. Any scores that were out-of-bounds of the original range were set to the bound (e.g. a corrected proportion of -5% was set to 0%). Critically, this approach ensures that the correction applied to a score was never based on the error of that score in the first place, such that any subsequent comparisons of expert and user scores were still valid.

After correcting the intensity and proportion metrics we calculated an H-score for each core as described in McCarty *et al* (1985)[23]. We calculated this semi-quantitative score by multiplying the proportion of cancer cells stained by the marker (0 to 100%) by the average intensity of the staining (0 to 3). The H-score is therefore between 0 and 300, with 0

indicating no cancer cells positive for the marker, and 300 indicating all cancer cells positive with maximum intensity.

Comparison of user and expert scores

We calculated Spearman's rank correlation between expert and crowdsourced scores for the H-score and proportion of cancer cells stained. For intensity of staining we used quadratic-weighted kappa[24].

Associations with clinical outcomes

We used Kaplan-Meier curves and Cox proportional hazards models to estimate the associations between crowdsourced scores and disease-specific survival (DSS, time from treatment to death due to bladder cancer). For each marker, we examined the associations of the proportion of cancer cells staining positive, of the intensity of staining and of the H-score with DSS. Associations between H-score and survival were assessed using the numerical value of the H-score and quartiles for the H-scores. Quartiles were calculated on the combined dataset of all cohorts to ensure comparability of estimates for different cohorts. In the 2006-9 radiotherapy cohort, a significant number (>2/3rds) of cores were unusable due to diathermy artefacts within the tissue, created at the time of transurethral resection of the bladder tumour, therefore patient scores were unlikely to be representative, hence the 2006-9 cohort was not included in further analysis. The analysis was done separately in the cystectomy cohort and the two radiotherapy cohorts (1995-9 and 2002-5). The proportional hazards assumption was assessed by examining scaled Schoenfeld residuals. The main analyses were done using the aggregated scores. We carried out the analysis with and without adjustment for age, T stage, N stage, grade, sex and hydronephrosis. As a sensitivity analysis we repeated the main analyses with all observations of crowdsourced scores (from all individuals who used the app), taking into account that there are multiple observations per patient. We also investigated the associations of each marker with DSS in

subgroups defined by low/high CK5 and CK20 expression, in the cystectomy cohort and in the combined radiotherapy cohort (1995–9 and 2002–5).

Results

Public engagement

The game went live on 9th October 2014 and data reported here are those deposited up to 28th September 2016. Data were collected over 721 days, with 148,349 app downloads. The total number of classifications was 4,744,339 (excluding lung and test cases), and the mean number of classifications per day was 6,580. The average time per classification was approximately 15 seconds, an estimate based on the app analytics data. The total time contributed (# classifications * time per classification) - excluding the gaming element of the app - was approximately 20,000 hours.

As typically observed for projects of this type, the rate of classifications altered markedly over the two years of play, increasing in response to marketing activity before returning to a low level "baseline" of activity. Marketing activity included paid-for social media and television advertising as well as spontaneous news coverage and celebrity endorsements (Figure 2). Activity level dropped in response to any technical issues preventing normal game function. In-game messaging or "push notifications" were employed to improve classification rate following the resolution of such issues. Generally, the classification rate lowered over time with over half of all classifications being scored in the first 4 months.

Improving efficiency of crowdsourced scoring

In early 2015, there was concern that users were finding it difficult to score the outer squares of a TMA core, which often contained only a handful of cells and therefore lacked the tissue structure that often helps distinguish cancer from non-cancer tissue. An analysis was performed using an MRE11 test set to assess the effect of scoring only the central 16 squares compared to all 36 squares. This revealed that these mostly empty segments could be discarded from analysis without detrimental effect on accuracy. Additionally, an interim analysis was performed in August 2015 to calculate accuracy of scoring as a function of the number of ratings per image. This was to see whether the datasets could be processed

more quickly by reducing the number of raters from 25 per segment (Figure 3A). Based on this figure it was decided that obtaining more than 5 ratings per segment, i.e. 80 ratings across the 16 segments of a core, would yield minimal additional accuracy. For example, with 80 raters per core the p21 stain is scored with an accuracy of 0.85. Having 1000 raters per core would yield an accuracy of 0.89. The trade-off, then, is to sacrifice 0.04 in accuracy to be able to analyse 12 times more stain types. Hence, we decided to use these additional ratings to analyse additional stains.

Comparison of crowdsourced scores to expert scores

Proportion of cancer cells stained, and intensity of staining

Although the H-score is the primary outcome of immunohistochemical analysis, examining citizen scientists' accuracy on the proportion of cancer cells stained and the intensity of staining is instructive. We observed a wide range of correlations for both proportion and intensity, from 0.17 to 0.87 for proportion and 0.19 to 0.86 for intensity (Figure S2 and Table 1). For example, estimating the proportion of cancer cells stained for TIP60 was difficult (Spearman correlation of 0.17 between crowdsourced and expert estimates), whereas the intensity of staining was poorly estimated for Ki67 (quadratic-weighted Kappa of 0.19 between crowdsourced and expert estimates). Other markers seemed considerably easier to score for the public, such as CK20 which was scored at an accuracy of >0.8 for both intensity of staining and proportion of cancer cells stained.

H-score accuracy

The H-score is a combination of the intensity and proportion estimates, and we calculated the correlation in H-score between expert and crowdsourced estimates (Figure 3B and Figure S3). There was no clear correlation with the time of the sample set entering into the game (Table S3). The correlation between crowdsourced and expert H-scores varied from

0.65 to 0.92 across the markers tested here, with six of 10 correlation coefficients at least 0.80.

Associations between marker scores and disease-specific survival

Having established that crowdsourcing can yield reasonably accurate classifications of IHC scores, we then moved to see if any of these scores predicted disease-specific survival. Details of cohorts are shown in Tables S4 and S5. We fitted univariable Cox proportional hazards models for each of the stains. Statistically significant associations between H-score and DSS were found for MRE11, CK20, p21, 53BP1, p53 and Ki67 IHC (Table S6). Due to multiple testing some significant associations may be due to chance. However, MRE11 and CK20 displayed consistent relationships between IHC and DSS.

Similar to previously reported findings[16,21], high MRE11 levels were found to be significantly associated with DSS in the radiotherapy cohorts but not the cystectomy cohort. Significance was observed in both the 1995-9 and 2002-5 radiotherapy cohorts when comparing the 1st quartile of H-scores to the 4th quartile (Table S6). Furthermore, when using a numeric H-score, rather than comparing quartiles, there was a significant association between MRE11 staining and DSS in the 1995-9 cohort (HR per unit increase in H-score 0.991, 95%CI: 0.986-0.997, p=0.004) and borderline-significant association in the 2002-5 cohort (HR 0.994, 95% CI: 0.987-1.000, p=0.060, Table S6, Figures 4 and 5). High MRE11 expression (above its median) was not significantly associated with DSS in the cystectomy cohort. It was significantly associated with a lower risk of death due to bladder cancer in the radiotherapy 1995-9 cohort (HR for high vs low 0.30, 95% CI: 0.13-0.69, p=0.004), but not in the 2002-5 radiotherapy cohort, although in the same direction.

For CK20 staining, H-score was significantly associated with DSS in the cystectomy and 1995-9 RT with a hazard ratio per unit increase for the 2002-5 cohort of 0.998, 95% CI: 0.994-1.001, p=0.20. CK20 levels above the median H-score were associated with improved

survival in the cystectomy (HR: 0.454, 95% CI: 0.227-0.909, p=0.026) and the 1995-9 (HR: 0.292, 95% CI: 0.134-0.638, p=0.002, n=65) cohorts (Figures 4, 5 and S4).

Multivariable Cox proportional hazards models were fitted on the 1995-9 and 2002-5 radiotherapy cohorts for MRE11 and CK20 stains using age, T stage, N stage, grade, sex and hydronephrosis as covariables. The cystectomy cohort was excluded from multivariable analysis due to missing data. Results from multivariable analysis were in accordance with univariable analysis. In multivariable analysis for MRE11 staining in the 1995-9 cohort (n=61) significant associations with DSS were identified for MRE11 H-score for unit increase (HR: 0.992, 95% CI: 0.984-1.000, p=0.05) and hydronephrosis (HR: 3.101, 95% CI: 1.154-8.335, p=0.02). In the 2002-5 cohort (n=73) only MRE11 H-score was significantly associated with DSS (HR: 0.990, 95% CI: 0.981-0.998, p=0.02). In multivariable analysis of CK20 staining in the 1995-9 cohort (n=59), a significant association was observed for CK20 H-score and DSS (HR: 0.993, 95% CI: 0.990-1.000, p=0.02). In the 2002-5 cohort (n=74) no significant association was found for CK20 or any of the variables analysed.

In the combined radiotherapy cohort, in the group with low CK5 expression (less than its median) CK20 and MRE11 were associated with a lower risk of bladder cancer death (HR per unit increase 0.994, 95% CI: 0.990-0.998, p=0.003, and 0.991 95% CI: 0.985-0.996, p=0.0007, respectively). In the high CK5 subgroup the associations were similar in direction and magnitude but only borderline significant. In the high CK20 subgroup, 53BP1 was associated with a lower risk of bladder cancer death (HR: 0.992, 95% CI:0 .985=0.999), p=0.02). MRE11 was also associated with a lower risk (HR: 0.989, 95% CI :0.981-0.997, p=0.007). None of the markers were significantly associated with outcome in the low CK20 group.

Discussion

Reverse the Odds was a novel approach aimed at improving the speed of IHC scoring. Mobile gaming technology was combined with crowdsourcing to bring citizen science to a wider user-base than other projects such as Cancer Research UK's *Cell Slider* and *Trailblazer*. We observed moderate to high agreement between crowdsourced and expert scores, and crowdsourced scores successfully identified at least two markers as predictive of survival in bladder cancer.

As implemented in this study, our approach did not increase the speed of IHC scoring, as initially anticipated. It took just under 2 years to score 16 IHC markers, far longer than it would have taken researchers. However, the use of crowdsourcing embedded in mobile phone technology is in its infancy, and throughout the course of RTO lessons were learnt that would speed up analysis more than 10-fold for future projects.

In early 2015, there was concern that users were finding it difficult to score the outer squares of a TMA core, which often contained only a handful of cells and therefore lacked the tissue structure that often helps distinguish cancer from non-cancer tissue. Reducing the number of segments to be scored for each core to the 16 central segments did not affect the accuracy of core scoring and increased the speed at which a core could be scored by the public. Furthermore, accurate results could be achieved with fewer users than initially thought. With these adjustments, it is estimated that RTO could have analysed all 11 bladder cancer stains in the first 2 weeks after release of the game, which would amount to a significant increase in scoring efficiency over traditional scoring methods.

One problem identified in RTO was the drop-off in user participation over time. Indeed, half of all image analysis was conducted in the first 4 months of the game's release (Figure 2). This pattern has been observed in other crowdsourcing ventures such as *Galaxy Zoo, Milky Way Project, Fraxinus, EteRNA, Foldit* and *Phylo*. It is often found that a small group of dedicated individuals contribute the bulk of classifications, with the majority of users only

contributing transiently, and with some registered users never actually participating[9,11,25]. To exploit the high uptake and number of analyses conducted in crowdsourcing applications after initial release, good systems need to be in place upfront, and it is important that a robust user engagement plan for the promotion of a project is in place from the outset.

A major issue in using crowdsourcing in molecular pathology studies is its reliability. In RTO we found the accuracy of public scoring to vary between immunostains. The lowest accuracy in public scores was seen in stains which were classed by experts as more difficult. Some stains were particularly challenging, eg. MRE11 and TIP60 which were scored with the lowest accuracy when compared to experts. MRE11 and TIP60 both show heterogeneous staining and can have weak non-specific staining in negative cells. Additionally, TIP60 IHC can also produce high levels of non-specific background staining in some cores. This reflects the potential need for improving contributor skills and ongoing quality assurance in such crowdsourcing projects. Previous work suggests that though tutorials can have short-term beneficial effects, a more critical predictor of accuracy is long-term engagement and training through experimentation and ongoing feedback[26,27].

Similar to *Cell Slider*[8], in RTO the public were asked to score an isolated segment of a TMA core. This step was taken to allow easy viewing of individual cells on a smartphone and eliminated the poor user experience of having to pinch-zoom. However, this user interface had the potential to limit the accuracy of scores generated by participants. The density of cancer cells can vary markedly across a TMA core. Using segmentation, the score from an area containing relatively few cancer cells is equal to the score derived from an area with a large number of cancer cells and has the potential to skew results, especially in cores where staining is heterogeneous. In this study we accepted this as additional noise, but future studies could account for such effects of segmentation through segment weighting. Furthermore, in viewing a whole tissue core, a scorer can get a ‘global’ view of the staining across the whole core and what level of staining most of the cancer cells exhibit, which can

aid in the accuracy of scoring. Another benefit of viewing a complete tissue core is that it can help in distinguishing cancer from normal tissue and infiltrative lymphocytes. Despite these drawbacks of using a small mobile screen, we considered the trend from desktop- and laptop-based internet use towards mobile use sufficiently strong to explore the viability of a smartphone-based solution.

In our analysis of the association of public scores with clinical outcome we observed that there were multiple unusable cores in our 2006-9 radiotherapy cohort. This was due to the TMA block having been used for previous studies, unlike the others. The result of previous sectioning from the TMA was a loss of total cores from the TMA, an increase in cores lacking tumour tissue and an increase in cores affected by diathermy artefacts arising from the transurethral bladder tumour resection. Corresponding to this, when MRE11 staining was re-optimized and new sections stained, there were only 524 cores suitable for analysis compared to the 831 originally stained cores cut from less depleted blocks. We therefore question the utility of TMAs for muscle-invasive bladder cancer, as it is likely that only the top sections of a TMA will give reliable, representative results.

In terms of the association of public scores with clinical outcome, MRE11 and CK20 staining showed the strongest associations with DSS. This is an encouraging result as these two proteins have been linked to DSS in MIBC previously.

MRE11 IHC has been directly reported to associate to DSS in MIBC by two independent research groups[16,21]. Both studies identified high levels of MRE11 (greater than the 1st quartile of the data) to be associated with improved DSS in MIBC following radiotherapy-based treatment but with no association with outcome following cystectomy. While the results of this present analysis only found MRE11 levels above the 4th quartile to be significantly associated with improved DSS, the public scores were accurate enough to at least identify MRE11 as a candidate marker for further investigation.

CK20 has been used to identify the luminal subtype of MIBC. Luminal MIBC is associated with better DSS compared to basal MIBC[12] and hence CK20 may be a potential prognostic biomarker for MIBC. In this study, the crowdsourced CK20 scores identified a significant association between higher levels of CK20 and improved patient outcome in both the cystectomy and the 1995-1999 radiotherapy cohort in keeping with CK20 being a prognostic marker for MIBC. In contrast, crowdsourced scores for the basal marker CK5/6 were not associated with DSS in this analysis, despite good agreement between experts and crowdsourced scores. This is surprising given that high levels of CK5/6 being previously associated with poor DSS[28,29,30].

In this study, we have gained insight into the potential advantages and disadvantages of using crowdsourcing for the analysis of molecular pathology studies. A major advantage of using crowdsourcing to analyse IHC data is the potential time a well-planned and optimised method could save for skilled researchers. For crowdsourcing, researchers would only be required to score a small subset of data to generate tutorial images and 10% comparison data, thus freeing up time for other work. Although RTO did not achieve improvements in scoring efficiency, a number of steps could have been taken to dramatically speed up analysis. First, all datasets can be prepared in advance (preprocessing, segmentation, colour inversion, and hosting) to rapidly switch to a new dataset once a previous set is completed. Such completions happen in sometimes unpredictable bursts e.g. due to media coverage. Second, prior to any media launch every effort should be made to optimise the number of ratings that are necessary per sample. Many ratings happen upon initial launch, and these are in effect wasted if too many ratings are collected per sample. In RTO, prior optimisation of number of raters as outlined in the results would have seen all datasets analysed within the first 14 days of launch. Third, a community should be fostered to encourage learning and continued engagement, dramatically increasing retention of users obtained through marketing. Fourth, microtasks could be distributed to users based on their ability as assessed through scoring expert-scored samples. If the lessons learnt from RTO

and *Trailblazer* are applied to future projects, crowdsourcing has the potential to accurately screen IHC data and greatly increase the speed of biomarker discovery from large IHC data sets.

References

1. Theodosiou Z, Kasampalidis IN, Livanos G, Zervakis M, Pitas I, Lyroudia K (2007) Automated analysis of FISH and immunohistochemistry images: A review. *Cytom Part A* **71A**: 439–450.

2. Di Cataldo S, Ficarra E, Macii E (2012) Computer-aided techniques for chromogenic immunohistochemistry: Status and directions. *Comput Biol Med* 2012; **42**: 1012–1025.

3. Levenson RM, Krupinski EA, Navarro VM, Wasserman EA, Mgode G, Jubitana M Pigeons (Columba livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images. *PLoS One* 2015; **10**: e0141357.

4. Lintott CJ, Schawinski K, Slosar A, Land K, Bamford S, Thomas D, *et al.* Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Mon Not R Astron Soc* 2008: **389**: 1179–1189.

5. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, *et al.* Predicting protein structures with a multiplayer online game. *Nature* 2010; **466**: 756-60.

6. Kawrykow A, Roumanis G, Kam A, Kwak D, Leung C, Zarour E, *et al.* Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. *PLoS One* 2012; **7**: e31362.

7. Lee J, Kladwang W, Lee M, Cantu D, Azizyan M, Kim H, *et al.* RNA design rules from a massive open laboratory. *Proc Natl Acad Sci U S A* 2014; **111**: 2122–2127.

8. Candido dos Reis FJ, Lynn S, Ali HR, Eccles D, Hanby A, Provenzano E, *et al.* Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer. *EBioMedicine* 2015; **2**: 681–689.

9. Rallapalli G, Fraxinus Players F, Saunders DG, Yoshida K, Edwards A, Lugo CA *et al.* Lessons from Fraxinus, a crowd-sourced citizen science game in genomics. *Elife* 2015; **4**: e07460.

10. Lawson J, Robinson-Vyas RJ, McQuillan JP, Paterson A, Christie S, Kidza-Griffiths M, *et al.* Crowdsourcing for translational research: analysis of biomarker expression using cancer microarrays. *Br J Cancer* 2017; **116**: 237–245.

11. Curtis V. .Online citizen science games: Opportunities for the biological sciences. *Appl Transl Genomics* 2014; **3**: 90–94.

12. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J *et al*. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell* 2014; **25**: 152–65.

13. Wu CS, Pollack A, Czerniak B, Chyle V, Zagars GK, Dinney CP, *et al.* Prognostic value of p53 in muscle-invasive bladder cancer treated with preoperative radiotherapy. *Urology* 1996; **47**: 305–310.

14. Qureshi KN, Griffiths TR, Robinson MC, Marsh C, Roberts JT, Lunec J, *et al.* Combined p21WAF1/CIP1 and p53 overexpression predict improved survival in muscle-invasive bladder cancer treated by radical radiotherapy. *Int J Radiat Oncol Biol Phys* 2001; **51**: 1234–1240.

15. Malats N, Bustos A, Nascimento CM, Fernandez F, Rivas M, Puente D, *et al.* P53 as a prognostic marker for bladder cancer: a meta-analysis and review. *Lancet Oncol* 2005; **6**: 678–686.

16. Laurberg JR, Brems-Eskildsen AS, Nordentoft I, Fristrup N, Schepeler T, Ulhøi BP, *et al.* Expression of TIP60 (tat-interactive protein) and MRE11 (meiotic recombination 11 homolog) predict treatment-specific outcome of localised invasive bladder cancer. *BJU Int* 2012; **110**: E1228–E1236.

17. Tanabe K, Yoshida S, Koga F, Inoue M, Kobayashi S, Ishioka J, *et al.* High Ki-67 Expression Predicts Favorable Survival in Muscle-Invasive Bladder Cancer Patients Treated With Chemoradiation-Based Bladder-Sparing Protocol. *Clin Genitourin Cancer* 2015; **13**:

e243–e251.

18. Tang K, Wang C, Chen Z, Xu H, Ye Z. Clinicopathologic and prognostic significance of p21 (Cip1/Waf1) expression in bladder cancer. *Int J Clin Exp Pathol* 2015; **8**: 4999–5007.

19. Tian Y, Ma Z, Chen Z, Li M, Wu Z, Hong M, *et al.* Clinicopathological and Prognostic Value of Ki-67 Expression in Bladder Cancer: A Systematic Review and Meta-Analysis. *PLoS One* 2016; **11**: e0158891.

20. Wang L, Zhou M, Feng C, Gao P, Ding G, Zhou Z, *et al.* Prognostic value of Ki67 and p63 expressions in bladder cancer patients who underwent radical cystectomy. *Int Urol Nephrol* 2016; **48**: 495–501.

21. Choudhury A, Nelson LD, Teo MTW, Chilka S, Bhattarai S, Johnston CF, *et al.* MRE11 expression is predictive of cause-specific survival following radical radiotherapy for muscle-invasive bladder cancer. *Cancer Res* 2010; **70**: 7017–7026.

22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011; **12**: 2825–2830.

23. McCarty KS, Miller LS, Cox EB, Konrath J, McCarty KS. Estrogen receptor analyses. Correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. *Arch Pathol Lab Med* 1985; **109**: 716–721.

24. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Phys Ther* 2005; **85**: 257-268.

25. Ponciano L, Brasileiro F, Simpson R, Smith A Volunteers' Engagement in Human Computation for Astronomy Projects. *Comput Sci Eng* 2014; **16:** 52–59.

26. Andersen E, O'Rourke E, Liu YE, Snider R, Lowdermilk J, Truong D, *et al.* The impact of tutorials on games of varying complexity. Proceedings of the SIGCHI Conference on Human

Factors in Computing Systems 2012; https://dl.acm.org/citation.cfm?id=2207687, accessed 30 May 2018, (2012).

27. Singh A, Ahsan A, Blanchette M, Waldispühl J. Lessons from an Online Massive Genomics Computer Game. Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing 2017; https://www.aaai.org/Library/HCOMP/hcomp17contents.php, accessed 30 May 2018, (2017).

28. Dadhania V, Zhang M, Zhang L, Bondaruk J, Majewski T, Siefker-Radtke A, *et al.* Meta-Analysis of the Luminal and Basal Subtypes of Bladder Cancer and the Identification of Signature Immunohistochemical Markers for Clinical Use. *EBioMedicine* 2016; **12**: 105–117.

29. Hayashi T, Sentani K, Kakumoto S, Oo HZ, Sakamoto N, Mutaguchi K, *et al.* Prognostic impact of immunohistochemical classification of bladder cancer according to luminal (Uroplakin III) and basal (Cytokeratin 5/6) markers. *Eur Urol Suppl* 2017;**16:** e681.

30. Zhang R, Chen H, Xia J, Shi O, Cao M, Jin D, *et al.* The pathological and clinical response of the luminal and basal subtypes of muscle-invasive bladder cancer to neoadjuvant cisplatin-based chemotherapy and radical cystectomy depend on the immunohistochemical classification system. Eur Urol Suppl 2017; **16**: e303–e304.

Acknowledgements

The game has been a joint collaboration between Cancer Research UK, Channel 4, Maverick Television, Chunk and Zooniverse. Only Cancer Research UK had access to the data and had control over the decision to publish. We would like to thank Dr Judith Nicholson, Dr Martin Kerr, Dr Eva McGrowder and Sarah Jevons for help with downloading and labelling individual TMA cores. We thank the rest of the Cancer Research UK Citizen science team: Josh Lee, Hannah Keartland, Rebecca White, Andy Patterson, Amy Garcia, Mary Cooke, Amber Holmes. We acknowledge funding from Cancer Research UK, Cancer Research UK programme grant C5255/A15935 (AEK) and the research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) (Molecular Diagnostics Theme / Multimodal Pathology Subtheme) (LB). AW was funded by an MRC studentship (MR/K501256/1).

Conflict of Interest

The authors declare no conflict of interest.

Table 1: Overview of markers, classifications and expert vs crowdsourcing scores.

Figure legends

Figure 1. Typical 0.6 um TMA core, stained with DAB and haematoxylin counterstain, and split into 6 x 6 grid. Upper left panel shows contents of red bound square colour transformed for use in the app by citizen scientists.

Figure 2. **Plot of user participation over time. (A)** Number of classifications per week. **(B)** Cumulative percentage of all classifications as a function of time.

Figure 3: A) Scatter plots for individual IHC stains ranked in order of H-score Spearman rho. X-axes represent the expert scores and y-axes the citizen score. Diagonal line represents a perfect score whereby the expert score is identical to the crowdsourced score; **B) The relationship between number of classifications and accuracy**. The y-axis represents the H-score Spearman rho between expert and crowdsourced scores, and the x-axis represents the number of classifications used for a core. The accuracy is estimated through bootstrapping with 1,000 samples. The error bars represent the bootstrapped 95% confidence interval (2.5 and 97.5 percentile of bootstrapped samples).

Figure 4: Kaplan Meier survival curves for disease-specific survival. MRE11, RAD50 and CK20 for 1995-9 radiotherapy cohort and CK20 cystectomy cohort.

Figure 5: Kaplan Meier survival curves for disease-specific survival. MRE11, RAD50 and CK20 for 2002-5 radiotherapy cohort.

Additional Information:

Ethics approval and consent to participate:

Ethical approval was obtained from London Bromley NRES (study 13/LO/0540), Leeds (East) Local Ethical Committee (studies 02/060 and 04/Q1206/62) and North West – Haydock Research Ethics Committee (study 14/NW/1033). Patients whose samples were collected from 2002 onwards gave informed consent for use of their pre-treatment biopsies.The study was performed in accordance with the Declaration of Helsinki.

Consent to publish:

Not applicable.

Availability of data and materials:

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Conflict of Interest:

The authors declare no conflict of interest.

Funding:

We acknowledge funding from Cancer Research UK, Cancer Research UK programme grant C5255/A15935 (AEK) and the research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) (Molecular Diagnostics Theme / Multimodal Pathology Subtheme) (LB). AW was funded by an MRC studentship (MR/K501256/1).

Authors' contributions:

Study conception and design: RR-V, LH, WH, AH, GJT, AEK.
Data production, analysis and interpretation: PS, AKW, SMcG, CK, RR-V, JPMcQ, SC, LH, JL, EH, WH, SB, LB, AEK.
Writing the manuscript: PS, AW, CK, RR-V, AEK.
All authors reviewed the manuscript and approved the final manuscript.

Acknowledgements:

The game has been a joint collaboration between Cancer Research UK, Channel 4, Maverick Television, Chunk and Zooniverse. Only Cancer Research UK had access to the data and had control over the decision to publish. We would like to thank Dr Judith Nicholson, Dr Martin Kerr, Dr Eva McGrowder and Sarah Jevons for help with downloading and labelling individual TMA cores. We thank the rest of the Cancer Research UK Citizen science team: Josh Lee, Hannah Keartland, Rebecca White, Andy Patterson, Amy Garcia, Mary Cooke, Amber Holmes.