Successful demand management in diagnostic immunology testing

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

OPEN ACCESS

Journal of
Clinical Pathology

Successful demand management in diagnostic immunology testing

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

TITLE PAGE

Successful demand management in diagnostic immunology testing

Kristen Lilly,[1] Nathan Proudlove[2] and Claire Bethune [1]
[ORCIDs: 0000-0003-0191-3701 0000-0002-1176-8088   0000-0002-5453-4927]

[1] Derriford Hospital, University Hospitals Plymouth NHS Foundation Trust, Plymouth, UK.
[2] Alliance Manchester Business School, University of Manchester, Manchester, UK

Correspondence to: Kristen Lilly
        Department of Immunology and Allergy
        Eden Unit, Level 7
        Derriford Hospital
        University Hospitals Plymouth NHS Trust
        Plymouth
        PL6 8DH
        kristenlilly@nhs.net

Keywords or phrases:
        Demand management, Diagnostic Overuse and Underuse, Immunology and Allergy Testing,
        Inappropriate tests

Word count (excluding title page, abstract, references, figures and tables): 5340

Ethics approval statements that refer to your institution:
        This Service Evaluation was approved and registered with the Clinical Audit Department of
        the University of Plymouth Hospitals NHS Trust in accordance with the Trust's
        policy. According to Health Research Authority (HRA) guidance, neither approval from an
        ethics committee nor informed consent is required for Service Evaluation reports.

Contributorship statement:
        KL and CB designed the project. KL obtained the data, which NP analysed. All authors
        interpreted the data and wrote and reviewed the drafts. CB is the guarantor of the
        manuscript.

1

Abstract [248 of 250 words]

Aims We investigated whether we could have a material and sustained impact on immunology test ordering by primary care clinicians by building evidence-based and explanatory algorithms into test ordering software.

Methods A service evaluation revealed cases of over-requesting of anti-nuclear antibody (ANA), allergen Specific Immunoglobulin E (IgE) and Total IgE tests, and under-requesting of Urine Protein Electrophoresis. We conducted a quality improvement programme to address this. We determined the most effective and efficient intervention would be to embed evidence- and advice-based decision-support algorithms in the ordering software. Consultation with GPs revealed lack of knowledge and confidence about testing, and an appetite for support. We iteratively designed and implemented algorithms for the four sets of tests for the primary care practices in our catchment and made them available to other hospital trusts in our region. The ordering system now contains links to advice sheets for clinicians and their patients and to an email address for queries to the lab.

Results We observe large (36% to 88%) reductions in testing activity (workload) for the over-requested tests and large (28% to 135%) increases for the under-requested test. We show that these changes are sustained. There have been no complaints from the clinicians and queries to the lab are now minimal (less than one per month on average).

Conclusions Embedding algorithms in the ordering software can be acceptable to clinicians and have a major and sustained impact on over- or underuse of tests. The algorithms can be replicated by other hospital trusts.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Diagnostic tests should be conducted when clinically useful but are frequently over- or under-requested.

WHAT THIS STUDY ADDS

⇒ We demonstrate that evidence- and advice-based decision-support algorithms built into test-ordering software can be acceptable to primary care physicians and can have a large and sustained impact on test ordering – decreasing either over- or underuse. Such algorithms can also be replicated by other hospital organisations.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Information and advice provided to physicians at the point of test request can help direct appropriate use of diagnostic tests, thereby improving investigation pathways and patient experience, and reducing waste of time and resources in primary care and testing laboratories. Test-order communication software is a helpful vehicle for making relevant information available and useful at the point of request.

INTRODUCTION

Ensuring that laboratory tests are requested only when clinically useful is a challenge for laboratory services and clinicians alike. Discussions with requesting clinicians in primary care highlighted the challenges they face in ensuring that they apply the most up-to-date testing guidelines in the time they have available with patients in clinic. This can lead to both over- and under-requesting of diagnostic tests, both of which can have a negative impact on patient care.[1] Estimates of the proportion of lab tests which are unnecessary range from 10-70%, depending on the nature of the test itself and the clinical situation.[2] There can also be considerable variation between the practices of ordering clincians.[3, 4] A meta-analysis has found that overutilisation rates of lab testing are highest for low-volume tests, and the rate of underutilisation (45% overall) tends to be much higher than overutilisation (21% overall), though underutilisation has been far-less studied.[5]

In this paper, we describe the development, implementation and impact of algorithms providing primary care clinicians with information to help them make decisions about test requesting while the patient is with them in the clinic. Such approaches have been successful in other clinical and geographical areas.[2, 6, 7] We described algorithms for four sets of diagnostic immunology tests (anti-nuclear antibodies [ANAs], allergen Specific Immunoglobulin E [IgE], Total IgE and Urine Protein Electrophoresis [UE]) and the impact of each on test activity volumes. This content of this paper is guided by the Standards for QUality Improvement Reporting Excellence (SQUIRE 2.0) framework.[8]

Immunology Testing

From a patient perspective, more testing does not always lead to a better clinical outcome. The statistics of setting reference ranges mean there can be a 5% false positive rate.[3] Tests interpreted outside the appropriate clinical context can lead to mismanagement; e.g., unnecessary food avoidance in the case of misinterpreted allergy tests, or anxiety if a positive test with low predictive value in a particular clinical context is misinterpreted. On the other hand, lack of familiarity with relatively rare conditions can mean clinicians may also underutilise some tests.[3]

Our discussions with clinicians highlighted the importance of providing timely and easily-accessible information indicating which tests were useful when making a new diagnosis, as well as making clear which tests are helpful when monitoring patients with a known diagnosis. The pressure from patients themselves to have particular tests was also raised by clinicians as an increasing challenge.

From the laboratory's perspective, quality improvement often focuses on efficiency in terms of reducing turnaround times,[9, 10] but consideration should also be given to the appropriateness of requests coming in 'through the door', in particular if screening requests or conducting inappropriate tests consumes staff and equipment capacity.

Demand Management Programme

As detailed in the next section, this work was driven by our knowledge that there was excessive ordering of some immunology tests (notably ANA, Specific IgE and Total IgE) and underuse of another (UE). Consideration of the options led us to conclude that the best approach would be to embed advice algorithms in our primary care order communications software (Integrated Clinical Environment, ICE). Engagement with referring clinicians revealed an appetite for advice, and we gained the cooperation of our in-hospital IT support, making this appear a feasible approach. Therefore:

- the aim of this study was to investigate whether we could have a material and sustained impact on immunology test ordering by primary care clinicians by building evidence-based and explanatory algorithms into the ordering software.

3

This paper reports the progressive development and implementation of such algorithms for these four sets of tests across our GP-referrer catchment, and the implementation of two of the algorithms in the catchments of two other hospital trusts in our region (one at each).

METHODOLOGY

The project had several phases:

- service evaluation
- consideration of alternative points of intervention
- consultation with referring GPs, focused on the ANA tests
- design of an algorithm for ANA test ordering
- implementation of the ANA algorithm and evaluation of the impact on the system
- replication of the approach for three other tests
- observation of the on-going impact at our trust and two others

Service evaluation

At University Hospitals Plymouth NHS Trust (UHP), service evaluation through snapshot data analysis and discussion with immunologist and rheumatologist colleagues revealed issues relating to test requests. Notable examples included:

- 10 successive ANA tests performed for the same patient within 12 months (tests could be re-ordered after 30 days)
- a request form containing orders for 59 Specific IgE tests for foods (all for the same patient) (following this up we discovered that the patient actually had *no* symptoms suggestive of food allergy)
- many queries by phone from GPs to various hospital clinicians and lab staff asking for interpretation of results (which we then determined had been requested outside the appropriate clinical context)

For a period during this time, we queried requests containing more than five Specific IgE tests with GPs, and often found that they did not understand when tests were appropriate or that additional tests had been added subsequently by nurses or phlebotomists. Using these query conversations as an opportunity to try to explain the indications for appropriate requests was (naturally) time-intensive and had limited reach, so (unsurprisingly) had no discernible impact on the scale of problem.

Points of Intervention

With a focus on inappropriate test ordering, there are several possibilities for intervention. Figure 1 shows a conceptual process map[11] of the testing process, highlighting the wastes[12, 13] generated by inappropriate requests. Demand management at the laboratory stage requires screening of requests once samples have reached the laboratory (Point B on Figure 1), seeking to avoid unnecessary lab work (Waste B). However, this relies on clinical details being recorded accurately by the requestor and, to be effective, requires experienced clinical/scientist review at the sample reception/booking-in stage. Even if adequate clinical information *is* available at this point to reject inappropriate requests, significant resources will already have been deployed requesting and taking the patients' blood, transporting it to the laboratory and booking the sample in at reception (Waste A).

[FIGURE 1 ABOUT HERE]

4

The outcome from an inappropriate test request, whether rejected by the lab or conducted and producing a positive or negative result, is likely to generate an additional appointment between the clinician and patient (Waste C) to discuss 'what's next?' (a change in investigation strategy or an [inappropriate] treatment or management strategy). From a patient's perspective, a test discussed with the clinician in a clinic but then rejected at Point B (after the sample has been taken) may undermine confidence in the whole diagnostic process.

To further illustrate our choices of intervention, Figure 2 shows a driver diagram[14, 15] (also known as an action effect diagram[16]) of our hypothesised causal logic and our options for intervention.

[FIGURE 2 ABOUT HERE]

It shows the three options we considered:

- Option 1: continue to make use of opportunities to discuss appropriateness of requests with GPs (when they make enquiries and through occasional follow-up query calls from us). Such education should impact the whole system, but has limited reach (especially for increasing requests for underutilised tests as omission leads to less contact with us) and is time consuming.
- Option 2: introduce screening (gatekeeping) at lab reception to reject inappropriate test requests received, Point B on Figure 1. This would not (much) reduce the upstream and downstream wastes (Wastes A and C), have no impact on underutilisation of tests and consume more lab resource (staff time).
- Option 3: an algorithm at point of request (Point A) is the only option with clear potential to have material effect on the whole system.

Option 3, intervention at decision Point A (the clinician's decision to order a test) is the most efficient for all parties: patient (experience), clinician (time) and the wider system (expense), and is in accordance with operations management system-design principles[13, 17, 18] (see also online supplement Appendix 1) and with the NHS England clinically-led improvement programme Getting It Right First Time (GIRFT).[19] We identified the ICE software as a suitable vehicle for such an intervention.

If the clinician orders one or more *other* test (appropriately) at the same time, removing an inappropriate test would not yield the full waste-reduction benefits. However, avoiding taking and sending an additional sample would still result in saving tubes and forms, and some staff time at each stage.

Metrics

As shown in Figure 2, the key metric, our outcome metric, was lab activity (tests completed) for the targeted tests. At UHP we did not (and do not) have gatekeeping at lab reception to reject inappropriate tests, so the activity volume mirrors the requests (demand), but we will use the term 'activity' as it is the exact variable in our data.

We hoped to bring benefits to GPs – or (crucially) at least not make them very dissatisfied and so (potentially) reject our intervention – so we have classed this as a process metric to be assessed to check that our change is in the right direction. Balancing metrics seek to pick up factors external to our intervention which could confound our intended causal impact, such as general changes in laboratory workload.

Consultation with GPs

5

In order to understand the perspective of requesting clinicians, we and a UHP Trust IT specialist held a focus group at a referring practice with six GPs with a spectrum of clinical experience. We discussed the findings from our service evaluation and our proposal for point-of-request algorithms. At the end of the meeting there was a consensus from the GPs in the group that they did want the system to help guide which investigations to request, but wanted as few 'clicks' as possible, so that the benefit of using an evidence-based and explanatory algorithm would outweigh the disadvantage that it would take them longer to make each individual request.

To get view of practice prior to intervention, we sent a simple survey questionnaire to 40 GPs who had recently ordered ANA tests (which we had decided to tackle first) and received 26 responses (response rate 65%). (See online supplement Appendix 2 for the questions and responses.) The results showed that half the respondents were unsure or very unsure as to when it was clinically appropriate to request an ANA test and over 80% were unsure or very unsure regarding the appropriate repeat interval for ANA. The lack of confidence and appetite for increased advice and information at the point of request gave us confidence to proceed with developing algorithms.

Algorithm Design

We developed test-specific algorithms based on national guidance (see Table 1) and supporting information leaflets for clinicians and GPs. A requestor is guided through a set of prompts relating to the individual patient's clinical presentation, and is presented with consequent recommended decisions about condition diagnosis and management. The UHP pathology IT team implemented these in our ICE software.

Table 1: Algorithm Basis & Design

ANAs: two-pronged approach
1. The algorithm advises requestors of a *minimum repeat interval* of three years*[20]   • If a result within three years of the current request is on record for the patient it will pop up on screen   • A link to information describing indications for ANA testing is provided   2. Requestors are asked to identify the correct clinical context of the request from a drop-down list of clinical conditions for which ANA is helpful for diagnosis. If the requestor is not considering one of the listed conditions, they are advised an ANA request is not appropriate*[21]
Total IgEs:
Requestors are advised that Total IgE testing is not recommended for routine* first-line testing in primary care. Instead requestors are provided with: • A link to further information regarding indication for Total IgE testing and interpretation of previously positive results[22]   • Contact information for further discussion and the capacity to request if required (for example if the request has been made by the respiratory or allergy teams)
Specific IgEs: split into allergen groups by patient presentation and allergen type: inhalants, drugs, venom and food.
- Inhalants (e.g., house dust mite, pollen and animals):
Testing is advised and available *only if* management in primary care will be affected by the result* (e.g., testing for animal dander leading to avoidance)*

6

• Links provided to NICE guidance for management of rhinitis[23] and asthma[24] and allergy clinic referral requirements for patients remaining symptomatic after first-line treatment.

- Drug allergens:
Requestors are advised that a negative Specific IgE to drugs does not necessarily exclude drug allergy
• Link provided to NICE guidance for referral to allergy clinic[25]

- Venom allergens:
• If patient had anaphylaxis/systemic allergic reaction, requestor directed to an order set for Specific IgE to Bee Venom, Wasp Venom and Mast Cell Tryptase and advised to consider referral to the allergy clinic in line with NICE guidance[26]
• If symptoms suggest local reaction, the request is not approved* and a link is provided to information regarding management of large local reactions

- Food Allergens: Three-question approach to each individual allergen
• Requestors are guided through a number of questions including about the nature, timing and reproducibility of symptoms for each allergen under consideration. The request is only approved if responses are consistent with type one hypersensitivity[27]*
• If the request is rejected following the questions described above, a link is provided to information about food allergy diagnosis and other food related symptoms (e.g., food intolerance)

Urine Electrophoresis
Requestors were advised that in order to complete a myeloma screen, urine should be sent for protein electrophoresis**

*Where the algorithm leads to a request being rejected, links to further information (including leaflets suitable for patients and clinicians are provided). A link directly to an email inbox is also provided so that in specific clinical circumstances the requestor's requirements can be discussed with a clinical scientist.

** fairly recently changes to myeloma screening has included addition of serum free light chains if urine samples have not been received[28]

Implementation, Evaluation and Replication

We undertook an initial six-week pilot of each algorithm with the GP practices who refer to the UHP laboratory. To monitor GP reactions, after ANA implementation at UHP we repeated the survey exercise and monitored the number of queries received in the email box linked from the algorithm. These are discussed in the Results section.

The main evaluation is of long timeseries of the number of tests performed (activity) per month of each of the four test types. Although not directly impacted by the ANA algorithm, we also analysed the number of double-stranded DNA (dsDNA) and extractable nuclear antigen (ENA) tests as these are often added-on to positive ANA results. We hypothesised that reducing the total ANA requests may also reduce ENA and dsDNA testing.

Following favourable results, we shared the algorithms with other local trusts so they could consider their own implementations. We were subsequently able to obtain data from two of these other trusts ('Trust A' and 'Trust B') where they chose to implement an algorithm (ANA at Trust A and UE at Trust B) together with a dataset from Trust B on Specific IgE to which they did *not* implement algorithmic

demand management and so is useful for comparison. Implementations occurred between August 2017 and October 2020, and all datasets run from January/April 2016 to October/November/December 2021, so each has around N=70 datapoints (months of activity volumes). This gives us good pre-implementation (baseline regime) vs. post-implementation (algorithm regime) data periods to robustly assess the impact of the interventions and also span the main impact of COVID-19 (spring 2020). Table 2 contains descriptive statistics for the datasets, split into pre- vs. post-implementation periods. Each data point is monthly primary-care test activity (i.e. the number of tests conducted in the month).

We analysed these with run charts and statistical process control (SPC) charts, as conventionally used in quality improvement projects.[29, 30] Further analyses used parametric and non-parametric statistical tests.

8

Table 2: Monthly activity (monthly primary-care test activity): Descriptive statistics

Minimum etc monthly activity during a regime period), changes (Δ) and percentage changes (%Δ) between the two regimes; N = number of datapoints (months), SD = standard deviation. The Δ Mean figures are those used in the later t-testing, they exclude 5 datapoints (3 months of COVID-19 impact and 2 months of later 'catchup'). Shaded row: algorithm not implemented, split at the time point the Specific IgE was implemented at UHP for comparison.

Test	Referrer	Regime	Start of Data	- End of Data	Activity (tests per month)									
					N	Min	Median	Max	Δ Median	%Δ Median	Mean	SD	Δ Mean	%Δ Mean
ANA	UHP	1.Baseline	01/04/2016	01/07/2017	16	105	134	164			135.50	16.84		
		2.Algorithm	**01/08/2017**	01/10/2021	51	19	87	145	-47	**-35%**	86.22	19.36	-49.41	**-36%**
	Trust A	1.Baseline	01/01/2016	01/09/2020	57	71	226	285			216.86	38.19		
		2.Algorithm	**01/10/2020**	01/12/2021	15	60	102	144	-124	**-55%**	100.87	27.93	-127.65	**-57%**
DNA+ENA	UHP	1.Baseline	01/04/2016	01/07/2017	16	41	59	74			60.19	9.03		
		2.Algorithm	**01/08/2017**	01/11/2021	52	13	52	92	-7	**-12%**	52.08	15.79	-7.55	**-13%**
Total IgE	UHP	1.Baseline	01/01/2016	01/09/2017	21	16	25	35			24.67	5.31		
		2.Algorithm	**01/10/2017**	01/11/2021	50	0	2	8	-23	**-92%**	2.74	2.21	-21.80	**-88%**
Specific IgE	UHP	1.Baseline	01/04/2016	01/11/2018	32	26	64.5	133			71.03	23.56		
		2.Algorithm	**01/12/2018**	01/10/2021	35	3	21	36	-43.5	**-67%**	20.54	9.00	-49.20	**-69%**
	Trust B	1.Baseline	01/04/2016	01/11/2018	32	284	376	462			381.63	44.80		
		NOT implemented	01/12/2018	01/11/2021	36	150	463	688	87	**23%**	455.83	90.03	87.60	**23%**
Urine Electrophoresis	UHP	1.Baseline	01/04/2016	01/03/2018	24	56	75.5	109			77.83	10.91		
		2.Algorithm	**01/04/2018**	01/10/2021	43	37	98	139	22.5	**30%**	98.28	21.57	21.77	**28%**
	Trust B	1.Baseline	01/04/2016	01/10/2018	31	74	102	163			100.00	16.94		
		2.Algorithm	**01/11/2018**	01/11/2021	37	89	233	290	131	**128%**	227.27	38.68	134.63	**135%**

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

RESULTS

Quantitative Results – Graphical Analysis

Table 2 shows the impact on the median and mean activity, in absolute and percentage terms. Figure 3 shows the monthly data for ANA and (the associated) DNA + ENA tests, in run chart format, comparing the median activity pre- and post-implementation of the algorithm.[29]

[FIGURE 3 ABOUT HERE]

This shows the very material drop in ANA activity at UHP (upper left): the median has decreased by 35% (47 tests per month, Table 2), and this has been sustained. As we expected, the DNA and ENA activity (often add-ons) has also declined (by 12%). The bottom graph shows the replication at Trust A, where the ANA algorithm was implemented approximately three years after UHP. (The delay was mainly a protracted wait for an update to their IT system.) Trust A saw an even larger reduction (55%, 124 fewer tests per month). The impact of the algorithm (a sustain step-change) is very clear from the graphs.

The impact of the first COVID-19 lockdown (March 2020) can also be seen, with a dramatic dip in April 2020; the spike in activity one year later (around March/April 2021) is difficult to explain, but corresponds with the easing of 2021 lockdown measures and could be a catch-up of demand from a year before.

The results of implementing the Total IgE algorithm on activity at UHP, in Figure 4 upper left, shows an even more dramatic reduction in testing than we saw for ANA. Activity falls to almost zero, consistent with the algorithm. A year later (2018) we applied the algorithm to the allergen Specific IgEs requests, producing the change shown in the top-right graph, representing a reduction of 67% (around 44 test requests per month).

[FIGURE 4 ABOUT HERE]

As a comparison case, the bottom-right graph shows Specific IgE activity at Trust B, where the algorithm was *not* implemented. We have split the data at the time corresponding to introduction at UHP (12/2018) for clearer comparison. We see no corresponding *drop* in activity at the same time, giving us confidence that our reduction at UHP was not caused by some unknown external driver. In fact activity at Trust B is *higher* post 12/2018. A first reaction might be that demand has been displaced from UHP to Trust B. However, firstly, the increase at Trust B is twice the size of the drop at UHP and, secondly, there are no logistical or contractual mechanisms for GPs in our catchment to request from Trust B, particularly at the scale involved. The increase remains unexplained, but could be a result of a deliberate or inadvertent change in a screening approach at Trust B.) The volume of the activity at Trust B indicates the potential to reduce (inappropriate) activity if the algorithm were to be used.

Figure 5 shows the impact of implementation for UE, where we knew tests were being *under*-utilised as many requests for 'myeloma screen' were being submitted without a paired urine sample at a time when serum free light chains were not routinely available to GPs. At UHP there has been a material increase in activity, though perhaps not immediate and perhaps ongoing. This may be a result of increasing familiarity amongst GPs. As show, seven months later Trust B also adopted this algorithm, producing a more dramatic and step-change increase.

[FIGURE 5 ABOUT HERE]

These run charts are a straightforward format and medians are robust to extreme values. They show convincing results. The next stage in quality improvement analysis is often SPC charts. The corresponding SPC chart for each of the graphs above is shown in The Online Supplement Appendix 3.

Quantitative Results – Statistical Analysis

The graphs in Figures 3-5 show that all interventions (implementations of the algorithm) produced a material impact on the volume of orders for their corresponding test. As with much good service/quality improvement, the differences are large enough not to require the support of formal statistical testing, but for completeness we include some here.

Table 3: Robust statistical tests on monthly activity (monthly primary-care test activity)
based on the run charts; N = number of datapoints (months); shaded row: algorithm not implemented, split at the time point the Specific IgE was implemented at UHP for comparison.

Test	Site	Monthly Activity		N useful	runs signal	shift test		crossings test	
		Δ median	%Δ median			longest run	> critical	n crossings	< critical
ANA	UHP	-47	-35%	67	TRUE	44	9	11	26
	Trust A	-124	-55%	71	TRUE	26	9	16	28
DNA+ENA	UHP	-7	-12%	63	TRUE	10	9	23	25
Total IgE	UHP	-23	-92%	69	TRUE	55	9	7	27
Specific IgE	UHP	-43.5	-67%	67	TRUE	35	9	16	26
	Trust B	87	23%	66	TRUE	15	9	16	26
Urine	UHP	22.5	30%	67	TRUE	17	9	25	26
Electrophoresis	Trust B	131	128%	66	TRUE	19	9	13	26

Table 3 shows statistics based on the run charts and their medians, shown in Figures 3-5. Using medians makes the analysis robust to unusual (non-random) incidents in the data, such as the sharp (but short) drop due to the impact of the COVID-19 shock observed around April 2020 and some subsequent peaks (also short-lived) observed in some of the dataseries around 6- or 12-months later as lockdowns eased. We also include a pair of robust (nonparametric) statistical tests used with run charts: the shift and crossings tests.[29, 31-33] The absolute and percentage changes in the medians before vs. after implementation emphasise the material significance (impact).

The shift test is based on the (Wald–Wolfowitz) runs test. It compares the length of the longest sequence of datapoints on the same side of the median with what would be expected if the *whole* (unsplit) series were an random process with constant median (c.f., each datapoint being like the toss of a coin to determine whether it is above or below the median value established during the pre-implementation baseline). Typically-recommended thresholds for the critical value (longest expected sequence) range from 5 to 8. In our analyses we have untypically-long timeseries and use comparators calculated using the actual lengths of our series with a Type I error rate of 5% (α = .05), so that they are robust to series length as well as to the distribution of the data.[29, 31-33] N.b. in these analyses, datapoints on the median line are discounted, which is why in some cases the '*N useful*' in Table 3 may differ from the '*N*'s in Table 2. In all cases the test shows that our data have longest runs that are far longer than the critical values, so the datasets are non-random: something statistically significantly non-random happened.

11

The crossings test is similar, comparing the number of times a dataseries (the blue lines on Figures 3-5) cuts across the baseline median (the red dotted lines).[31, 32] This time, fewer than the expected minimum number of crossings signals non-random behaviour. Again, all are statistically significant.

The statistics also show that for Specific IgE at Trust B (the comparison site where the algorithm was *not* implemented) there has also been a non-random change, but in the *opposite* direction to that of the activity-reducing implementations of the algorithms.

Strictly, t-tests are less valid in this situation since we don't have two independent random samples from much larger populations, and they are less robust to very large or small values ('outliers') in data sets. They also lose the richness of the time-sequence by pooling data sets to just two time intervals (before vs. after). However, they are frequently used in analysis of change,[34] and we include them here in Table 4 for completeness. We show the changes in the means, and the percentage changes that these represent. These are very similar to the medians (Table 3) and also very statistically significant (against a usual p-value threshold of ≤ 0.05). The COVID-19 lockdown shock (grey vertical lines on the graphs) and the month either side, plus two-months of (possible) catchup compensatory peaks a year later, are excluded from these statistics and parametric tests.

Table 4: Parametric statistical tests on monthly activity
shaded row: algorithm not implemented, split at the time point the Specific IgE was implemented at UHP for comparison.

| Test | Site | Monthly Activity | | t-test of difference in means (Welch two-tailed) | | | |
		Δ mean	% Δ mean	t-stat	p-value	95% CI for change in mean	
ANA	UHP	**-49.41**	**-36%**	10.60	**<.001**	39.75	59.08
	Trust A	**-127.65**	**-57%**	15.96	**<.001**	110.97	144.32
DNA+ENA	UHP	**-7.55**	**-13%**	2.50	**.017**	1.45	13.64
Total IgE	UHP	**-21.80**	**-88%**	18.03	**<.001**	19.30	24.30
Specific IgE	UHP	**-49.20**	**-69%**	11.15	**<.001**	40.27	58.13
	Trust B	**87.60**	**23%**	-6.28	**<.001**	-115.58	-59.62
Urine	UHP	**21.77**	**28%**	-5.94	**<.001**	-29.10	-14.44
Electrophoresis	Trust B	**134.63**	**135%**	-25.06	**<.001**	-145.39	-123.86

Qualitative analysis – GP reactions

Following implementation of the ANA algorithm at UHP we repeated the survey of GPs who had recently requested ANAs. We had 18 responses. In supplementary material Appendix 2 the results are compared to the pre-implementation results. (They are not the same respondents prior and post – such a pairwise matching would have been impracticable, requiring them *all* to have ordered ANA tests again, and to respond, or else just selecting a small subset of responses with its own potential biasing.) We see increased confidence in when it is appropriate to order ANA (from 50% to 61% of respondents) and about the appropriate repeat interval (from 17% to 67%). More found ICE to be frequently clinically useful (from 69% to 88%) and useful for advice to patients (from 30% to 41%).

In the survey, we included provision for GPs to express concerns or dissatisfaction. There were no material comments about dissatisfaction (other than one about being slowed down by being able to accidentally click on a link and so open a page they had not meant to). There was no suggestion that

any GP had found the algorithm to have had an important negative impact on their requesting practice.

Prior to the algorithms, GPs would contact a range of people at the hospital with queries, and through a number of channels, so we had no way of monitoring all these. As noted under Table 1, the algorithms contain a link to an email box at the lab for queries. This prominent link makes it easy for them to contact us, and the unified channel enables us to see the volume of enquiries. Following implementation of the algorithms we have been receiving on average less than one enquiry per month.

DISCUSSION

Impact on test volumes

A systematic review of interventions to improve appropriateness of ordering by GPs (up to early 2014) found reductions of 1.2% to 60% were achieved, though with uncertain replicability.[35] More-recent studies include introduction of a test protocol for Lipid panels and haemoglobin A1c tests in a community-based clinic in the US resulting in a reduction of 50% in overuse[36], and pop-up reminders (about repeats, appropriateness and need for justification) in the capital region of Denmark resulting in a 25% reduction in vitamin D test requests from GPs.[37]

A review of ordering for in-hospital patients, published in 2017, found that: education had some impact on test ordering, but this was not sustained; feedback on comparative ordering practices amongst clinicians reduced frequent tests by around 11%, with another (small) study achieving 66% reduction in ionized calcium (but without much data on sustainability); and restriction on ordering within electronic systems reduced a set of common tests by 21%.[38] More recently, a decision-recommendation algorithm for calcium testing for inpatients in a hospital Brazil decreased ordering by 23%.[39]

Though we don't know appropriateness rates, the impacts we report in this paper on lab test activity (reflecting requesting rates) for our three overused tests (36% to 88% reductions) are at the high end of (or higher than) those reported in the literature. We also show evidence of the impact being sustained for a considerable period, along with some evidence of replicability at two other trusts. This demonstration of the effectiveness of algorithms embedded in test-requesting IT platform (ICE in this case) is a step towards addressing the problem of inappropriate over- or under-ordering of tests noted in the literature.[1, 2, 4, 5]

The changes in test activity we observed correspond to the dates of implementation of each algorithm at each trust and have been sustained, at UHP for over three years, and for nearly two years at Trust A. No corresponding reduction is observed in the dataset without the intervention (Trust B's allergen Specific IgE).

A possible confound to our results would be very substantial and sustained drops in requesting activity (or an increase in the case of UE) for some other reason(s), at or around the same times as the implementations of an algorithms. The replications across tests and sites, and the comparison with non-implementation for Specific IgE at Trust B, suggest this is not credible. As a further check, we have also looked at total UHP immunology lab workload for 2017/18 to 2021/22 as a balancing metric,[29] see supplementary material Appendix 4 (Figure A4.1). This shows a fairly steady, relatively small, increase (with a dip in 2021-22, which is almost certainly COVID-19 impact).

13

All our datasets are long and span the onset of the COVID-19 pandemic, in particular the (first, most-restrictive) UK 'lockdown'. This, of course, had a major effect on the testing process for patients, GPs and laboratories, as reported in the literature.[40-42] We can see from our data (Figures 3-5) that the effect was very short term for all tests covered in this paper. The grey vertical lines on the figures pick out April 2020. We can see this was the month most affected, with noticeable effects in the month each side; the material impact only covered the three months March-May 2020. After that we see activity return to the pre-March 2020 level (albeit commonly with a spike later that appears to be catch-up of the 'missing' demand displaced to post-lockdown). In fact our data series show faster and fuller recovery soon after the April nadir than for tests such as HbA1c analysed in other UK studies.[40, 41] This same pattern is also seen in our non-intervention (non-algorithm) data set (Specific IgE at Trust B - Figure 4 lower right). We argue, therefore that our results are not materially confounded by COVID-19 impact. We also removed these three 'COVID-19' months and two 'catch up' months a year later from our parametric statistics (Table 4) and SPC analysis (Appendix 3).

The run charts may also show some hints of the impact of the major blood-tube shortage, and consequent NHS-mandated restrictions on blood tests, at the end of August and first half of September 2021.[43] However, the graphs show the impact on the tests of interest here were, at most, minor and short-lived.

Point-of-request intervention

We implemented demand management at point-of-request (Point A on Figure 1) by developing algorithms to embed guidance for both clinicians and patients, recognising that some testing is driven by patient demand. We embedded links to short information leaflets as well as to national guidance in the requesting platform, so that the requestors could refer to them at the time of the consultation and, if desired, provide copies for patients. We also embedded an email link to unify channels for receiving queries from GPs.

A recent review of overordering of daily blood tests in hospitals adapted an effectiveness hierarchy for types of intervention to reduce inappropriate ordering.[44] Under this, ours is 'automation and computerization', a design-oriented intervention, and second in effectiveness only to *forced* restraint – i.e. vetting, disallowing or restricting ordering to limited time windows. Poorly-designed computerisation (e.g., poor arrangement of on-screen presentation of test order sets) can actively encourage poor clinician decisions by ordering staff; instead good design makes use of 'nudge' principles and defaults to encourage 'good' user behaviour.[7, 44] Further, computerised ordering can have large, wide and systemic impact to standardise practice; however, access to resources to modify IT systems can be a barrier.[44] We were fortunate to have good cooperation from our trust's Pathology IT team. The algorithms could also be a useful method to disseminate and implement new or updated local or national guidance on testing.

GP Satisfaction

A large survey of US GPs a decade ago found that, though they ordered tests from 31% of patient appointments, 15% expressed uncertainty about test ordering and 8% about test result interpretation; they welcomed IT decision support, but such systems were rarely available to them.[45] This covered a range of tests (including higher-frequency types). Our small-scale GP survey found higher levels of uncertainty about a low-frequency test (ANA) and a similar welcoming of IT decision support. Feedback from GPs through direct communication and our re-survey after the first implementation (ANA), suggested increased knowledge and confidence around test ordering and increased perceived

14

value of the ICE system. Subsequent queries from GPs were very infrequent (less than one per month on average).

Were initially concerned that GPs might object to the time that working through an algorithm would take (and the number of mouse clicks). On the other hand, these are relatively low-frequency tests for GPs, so not a large part of their workload and with consequently relatively-low familiarity with official guidance. Engaging with a focus group of GPs, we found initial resistance, but after informed discussion, they became comfortable with the idea and expressed an appetite for evidence-based support and advice.

Potential of the approach

In 2019, after our implementations at UHP, we surveyed the other acute trusts in the South West England region to understand whether any demand management procedures were in place for routine immunology-test requests from primary care. Of the 11 trusts that responded, seven (64%) had implemented demand management procedures for ANA requests and eight (73%) for allergen Specific IgE testing. However, *all* of these demand management procedures had been implemented at the laboratory stage (Point B in Figure 1), rather than the requesting stage (Point A), so incurring the cost of decision making time in the laboratory to screen samples received, together with the waste upstream (Waste A in Figure 1) and downstream (Waste C). The trusts not doing demand management at all also incur Waste B. We have demonstrated that our algorithmic approach can be applied across different trusts, so the potential is clear. We are now considering extending it to other tests.

Weaknesses

We did not attempt to directly assess the rate of over- or under-requesting of tests. We would have had an indication of the former if we had been previously been screening at lab reception. However, we knew from service evaluations that there were problems. The size of the impacts on activity we demonstrate in the data in this paper suggests the scale of prior over- or under-requesting. Further comparative datasets from other trusts' internal data systems, together with detail about any screening they did over the course of that data, would have been valuable but is very hard to obtain.

Whilst GPs have confirmed that they consider the algorithms worthwhile them taking the time to use, we were targeting what are (from a GP's point of view) low-frequency (fairly rare) tests, and so ones with which they tend to have low-familiarity. This approach may well not be acceptable to GPs for high-frequency tests, so a larger part of their workload.

The survey data regarding GP satisfaction was a relatively small sample (26 and 18 responses from 40 surveyed each time). The positive findings, though, are reinforced by the very low level of email queries to the laboratory regarding these tests, supporting their being general increased confidence amongst clinicians.

The GP ICE requesting system was a suitable platform in which to embed the algorithms. However, it is currently not possible to extend this to our in-hospital patient environment as a different requesting platform is used. To be effective, these algorithms rely on the request being made when the patient is present with the requestor to answer the questions posed. This is not always working practice (e.g., nurses asked to make requests on behalf of doctors).

CONCLUSIONS

15

This paper has demonstrated that algorithmic demand management can successfully guide requesting of immunology tests at the point of request. We have demonstrated that, as a result of implementation of the algorithms, the volume of testing activity (i.e., the volume of requests to our lab at UHP) was considerably reduced for previously over-requested tests (ANA, Specific IgE and Total IgE), avoiding waste in primary care, transport and the lab and supporting better GP and patient experience. The risk of misinterpretation of test results provided in an inappropriate clinical context is also reduced. This conforms with the GIRFT agenda, and contributes to creating capacity and reducing costs in the laboratory to free resources for other work such as the introduction of new assays. We have also demonstrated considerably increased activity (requests) for a previously under-requested test (UE). We made these changes without a negative reaction from GPs, in fact there are indications of increased confidence and knowledge and minimal need to contact the lab with queries. We have also shown that this approach is replicable at other trusts.

References

1. Hensher M, Canny B, Zimitat C, et al. Health care, overconsumption and uneconomic growth: A conceptual framework. *Social Science & Medicine* 2020;**266**:113420. doi:10.1016/j.socscimed.2020.113420

2. Beriault DR, Gilmour JA, Hicks LK. Overutilization in laboratory medicine: tackling the problem with quality improvement science. *Crit Rev Clin Lab Sci* 2021;**58**(6):430-46. doi:10.1080/10408363.2021.1893642

3. Naugler C, Wyonch R. What the doctor ordered: improving the use and value of laboratory testing. 2019. Canada: CD Howe Institute. www.cdhowe.org/public-policy-research/what-doctor-ordered-improving-use-and-value-laboratory-testing. (accessed 19 August 2022).

4. O'Sullivan JW, Albasri A, Nicholson BD, et al. Overtesting and undertesting in primary care: a systematic review and meta-analysis. *BMJ Open* 2018;**8**(2):e018557. doi:10.1136/bmjopen-2017-018557

5. Zhi M, Ding EL, Theisen-Toupal J, et al. The Landscape of Inappropriate Laboratory Testing: A 15-Year Meta-Analysis. *PLoS One* 2013;**8**(11):e78962. doi:10.1371/journal.pone.0078962

6. Taher J, Beriault DR, Yip D, et al. Reducing free thyroid hormone testing through multiple Plan-Do-Study-Act cycles. *Clin Biochem* 2020;**81**:41-46. doi:10.1016/j.clinbiochem.2020.05.004

7. Mathura P, Boettger C, Hagtvedt R, et al. Does admission order form design really matter? A reduction in urea blood test ordering. *BMJ Open Qual* 2021;**10**(3):e001330. doi:10.1136/bmjoq-2020-001330

8. Ogrinc G, Davies L, Goodman D, et al. SQUIRE 2.0 Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. *BMJ Qual Saf* 2016;**25**(12):986-92. doi:10.1136/bmjqs-2015-004411

9. White E, Proudlove N, Kallon D. Improving turnaround times for HLA-B*27 and HLA-B*57:01 gene testing: A Barts Health NHS Trust Quality Improvement project. *BMJ Open Qual* 2021;**10**(3):e001538. doi:10.1136/bmjoq-2021-001538

10. Li Y, Proudlove N. A quality improvement project to reduce the turnaround times of infectious disease markers reporting in an NHS stem cell department. *BMJ Open Qual* 2022;**11**(2):e001814. doi:10.1136/bmjoq-2022-001814

11. Trebble TM, Hansi N, Hydes T, et al. Process mapping the patient journey: an introduction. *Br Med J* 2010;**341**:394-97. doi:10.1136/bmj.c4078

12. Resar R, Griffin F, Kabcenell A, et al. Hospital Inpatient Waste Identification Tool. 2011. Cambridge, Massachusetts: : Institute for Healthcare Improvement. www.ihi.org/resources/Pages/IHIWhitePapers/HospitalInpatientWasteIDToolWhitePaper.aspx. (accessed 6 September 2020).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

13. Womack JP, Jones DT. *Lean thinking: banish waste and create wealth in your corporation*. revised & updated ed. London: Simon and Schuster, 2003.
14. Bennett B, Provost L. What's your theory? Driver diagram serves as tool for building and testing theories for improvement. *Quality Progress* 2015;**48**(7):36-43.
15. Langley GJ, Moen RD, Nolan KM, et al. *The improvement guide: A practical approach to enhancing organizational performance*. 2nd ed. San Francisco: Wiley, 2009.
16. Reed JE, McNicholas C, Woodcock T, et al. Designing quality improvement initiatives: the action effect method, a structured approach to identifying and articulating programme theory. *BMJ Qual Saf* 2014;**23**(12):1040-48. doi:10.1136/bmjqs-2014-003103
17. Boaden R, Harvey G, Moxham C, et al. Quality improvement: theory and practice in healthcare. 2008. Coventry: NHS Institute for Innovation and Improvement. www.england.nhs.uk/improvement-hub/wp-content/uploads/sites/44/2017/11/Quality-Improvement-Theory-and-Practice-in-Healthcare.pdf. (accessed 12 August 2021).
18. Seddon J. *Systems thinking in the public sector*. Axminster, UK: Triarchy Press, 2008.
19. GIRFT. The Getting It Right First Time programme. 2022. www.gettingitrightfirsttime.co.uk. (accessed 20 January 2022).
20. Arbuckle MR, McClain MT, Rubertone MV, et al. Development of Autoantibodies before the Clinical Onset of Systemic Lupus Erythematosus. *N Engl J Med* 2003;**349**(16):1526-33. doi:10.1056/NEJMoa021933
21. UHPT. ANA Requesting. n.d. www.plymouthhospitals.nhs.uk/eden-ana-requesting. (accessed 20 January 2022).
22. UHPT. IgE Testing. n.d. www.plymouthhospitals.nhs.uk/ige-testing. (accessed 20 January 2022).
23. NICE. Allergic rhinitis: Scenario: Management [Updated August 2021]. 2021. National Institute for Health and Care Excellence. https://cks.nice.org.uk/topics/allergic-rhinitis/management/management. (accessed 21 September 2021).
24. NICE. Asthma: diagnosis, monitoring and chronic asthma management (NG80) [Updated 22 March 2021]. 2017. National Institute for Health and Care Excellence. www.nice.org.uk/guidance/ng80. (accessed 21 July 2021).
25. NICE. Drug allergy: diagnosis and management (CG183). 2014. National Institute for Health and Care Excellence. www.nice.org.uk/guidance/cg183. (accessed 21 July 2021).
26. NICE. Anaphylaxis: assessment and referral after emergency treatment (CG134) [Updated 24 August 2020]. 2011. National Institute for Health and Care Excellence. www.nice.org.uk/guidance/cg134. (accessed 21 July 2021).
27. NICE. Food allergy: Quality standard (QS118). 2016. National Institute for Health and Care Excellence. www.nice.org.uk/guidance/qs118. (accessed 21 July 2021).
28. NICE. Myeloma: diagnosis and management (NG35) [Updated 25 October 2018]. 2016. National Institute for Health and Care Excellence. www.nice.org.uk/guidance/ng35. (accessed 21 July 2021).
29. Provost LP, Murray SK. *The health care data guide: learning for data improvement*. San Francisco, CA: Jossey-Bass, 2011.
30. Mohammed MA. Using statistical process control to improve the quality of health care. *Qual Saf Health Care* 2004;**13**(4):243-45.
31. Anhøj J, Olesen AV. Run charts revisited: a simulation study of run chart rules for detection of non-random variation in health care processes. *PLoS One* 2014;**9**(11):e113825. doi:10.1371/journal.pone.0113825
32. Anhøj J. Diagnostic Value of Run Chart Analysis: Using Likelihood Ratios to Compare Run Chart Rules on Simulated Data Series. *PLoS One* 2015;**10**(3):e0121349. doi:10.1371/journal.pone.0121349
33. Anhøj J, Wentzel-Larsen T. Sense and sensibility: on the diagnostic value of control chart rules for detection of shifts in time series data. *BMC Med Res Methodol* 2018;**18**(1):100. doi:10.1186/s12874-018-0564-0

17

34. Fretheim A, Tomic O. Statistical process control and interrupted time series: a golden opportunity for impact evaluation in quality improvement. *BMJ Qual Saf* 2015;**24**:748-52. doi:10.1136/bmjqs-2014-003756

35. Cadogan SL, Browne JP, Bradley CP, et al. The effectiveness of interventions to improve laboratory requesting patterns among primary care physicians: a systematic review. *Implement Sci* 2015;**10**(1):167. doi:10.1186/s13012-015-0356-4

36. Glauser E. Implementing an Evidence-Based Protocol to Reduce Inappropriate Laboratory Test Ordering. *The Journal for Nurse Practitioners* 2022 doi:10.1016/j.nurpra.2022.05.021

37. Munk JK, Bathum L, Jørgensen HL, et al. A compulsory pop-up form reduces the number of vitamin D requests from general practitioners by 25 percent. *Scand J Prim Health Care* 2020;**38**(3):308-14. doi:10.1080/02813432.2020.1794399

38. Eaton KP, Levy K, Soong C, et al. Evidence-Based Guidelines to Eliminate Repetitive Laboratory Testing. *JAMA Intern Med* 2017;**177**(12):1833-39. doi:10.1001/jamainternmed.2017.5152

39. Faria DK, Taniguchi LU, Fonseca LAM, et al. Improving serum calcium test ordering according to a decision algorithm. *J Clin Pathol* 2019;**72**(3):232-36. doi:10.1136/jclinpath-2018-205026

40. Kruger E, Banderker R, Erasmus R, et al. The impact of COVID-19 on routine patient care from a laboratory perspective. *S Afr Med J* 2020;**110**(12):1201-05. doi:10.7196/SAMJ.2020.v110i12.15294

41. Holland D, Heald AH, Stedman M, et al. Assessment of the effect of the COVID-19 pandemic on UK HbA1c testing: implications for diabetes management and diagnosis. *J Clin Pathol* 2021;**Online First**(13 October 2021) doi:10.1136/jclinpath-2021-207776

42. Durant TJS, Peaper DR, Ferguson D, et al. Impact of COVID-19 pandemic on laboratory utilization. *J Appl Lab Med* 2020;**5**(6):1194-205. doi:10.1093/jalm/jfaa121

43. Rimmer A. What has caused the NHS blood tube shortage, and how is it affecting doctors and patients? *BMJ* 2021;**374**:n2174. doi:10.1136/bmj.n2174

44. Jalbert R, Gob A, Chin-Yee I. Decreasing daily blood work in hospitals: What works and what doesn't. *Int J Lab Hematol* 2019;**41**(S1):151-61. doi:10.1111/ijlh.13015

45. Hickner J, Thompson PJ, Wilkinson T, et al. Primary care physicians' challenges in ordering clinical laboratory tests and interpreting results. *J Am Board Fam Med* 2014;**27**(2):268-74. doi:10.3122/jabfm.2014.02.130104
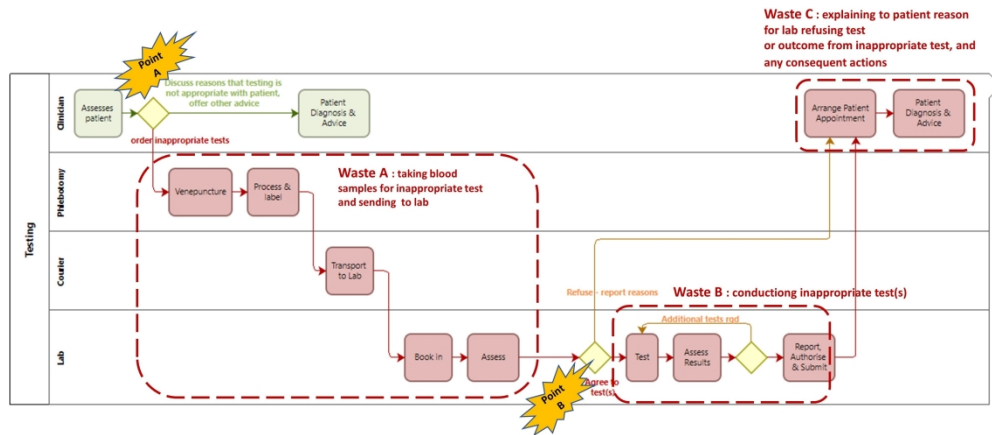
18

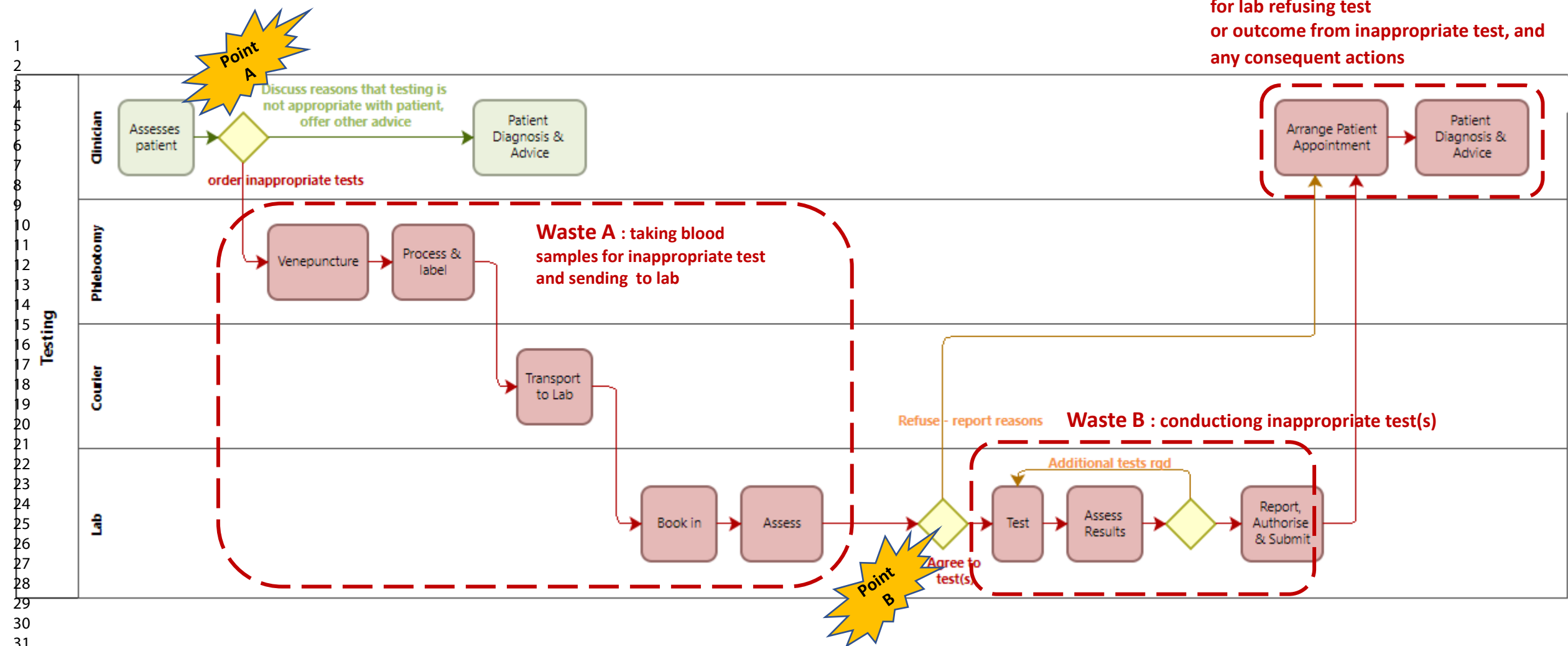Figure 1: Process map of inappropriate test ordering process, showing potential demand management intervention points Green = steps that are value to patient; Red = waste (non-value-adding or unnecessarily repeated steps); Orange line = necessary as a consequence of earlier waste

NOTE: for review we have added a 2nd supplementary pdf file containing copies of these main paper figures, which may be easier to read – Scholar one converts the 300dpi submission to low res when embedding them in the review document!

336x146mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
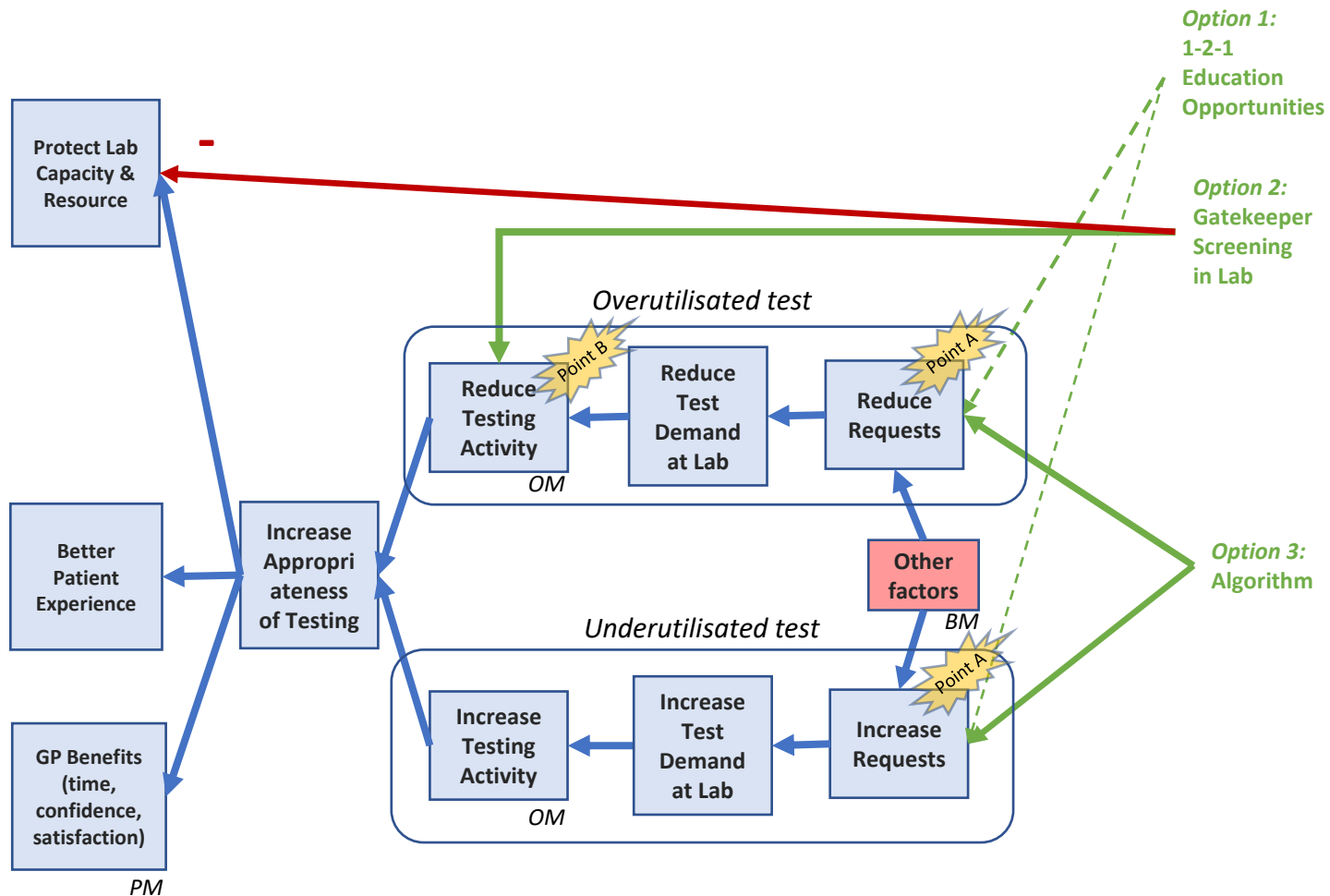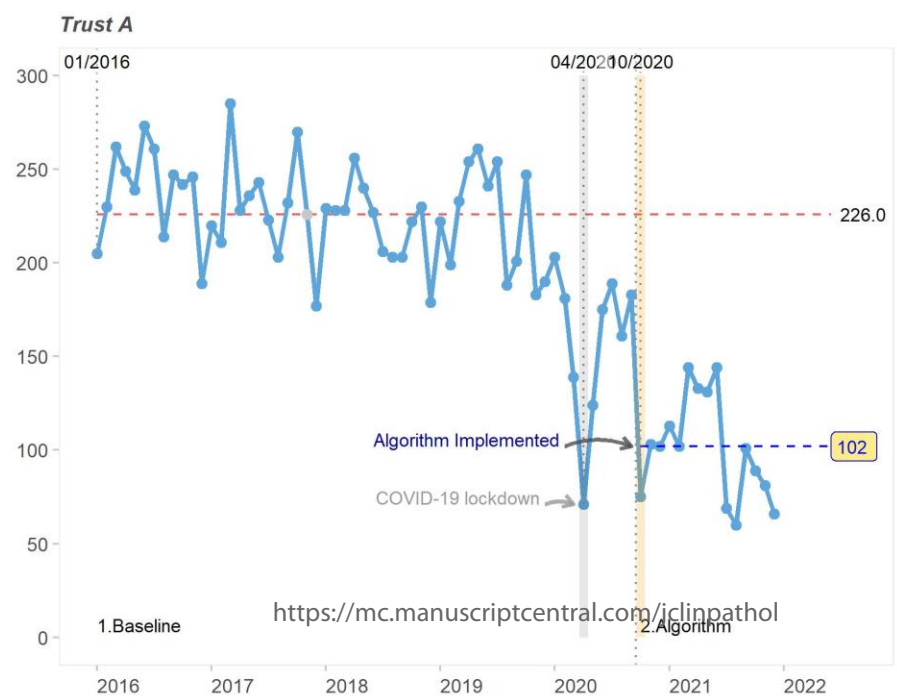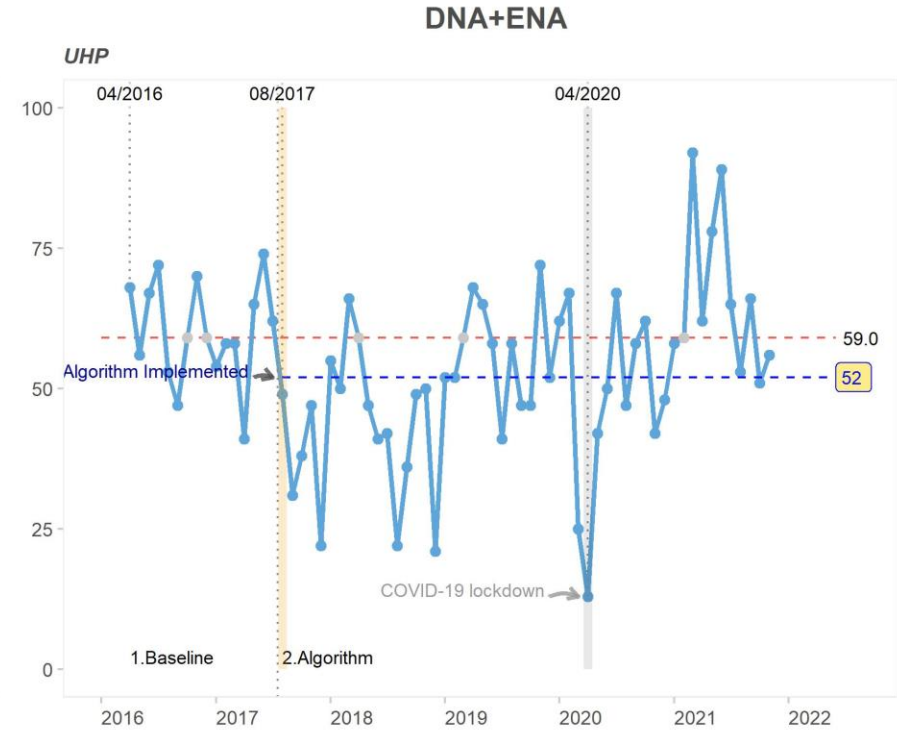47
48
49
50
51
52
53
54
55
56
57
58
59
60
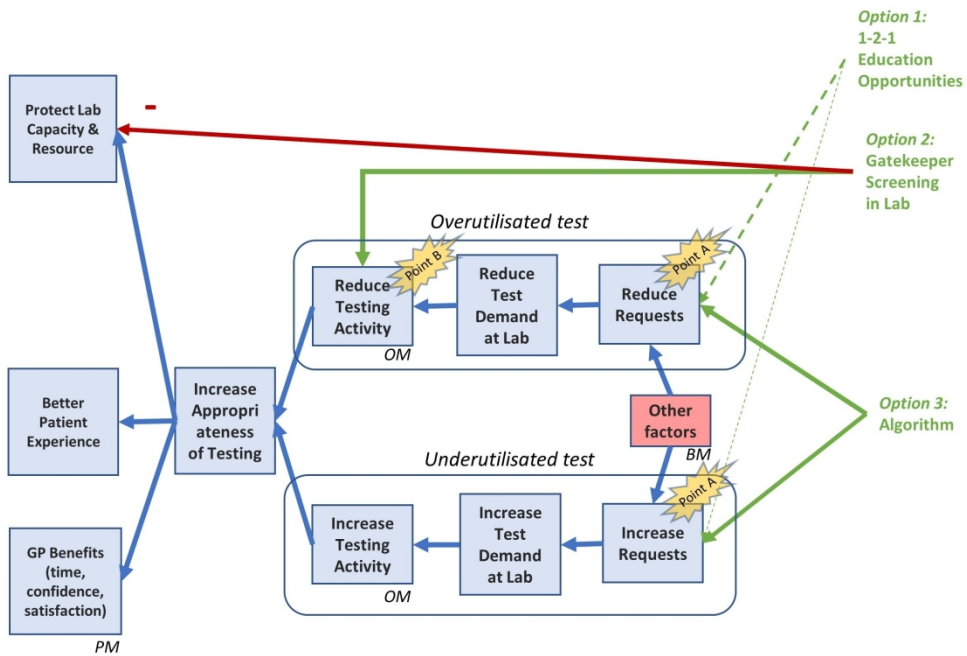
Figure 2: Driver Diagram showing hypothesised causal logic, options for intervention and points of measurement (OM = outcome metric, PM = process metric, BM = balancing metric)

197x131mm (300 x 300 DPI)
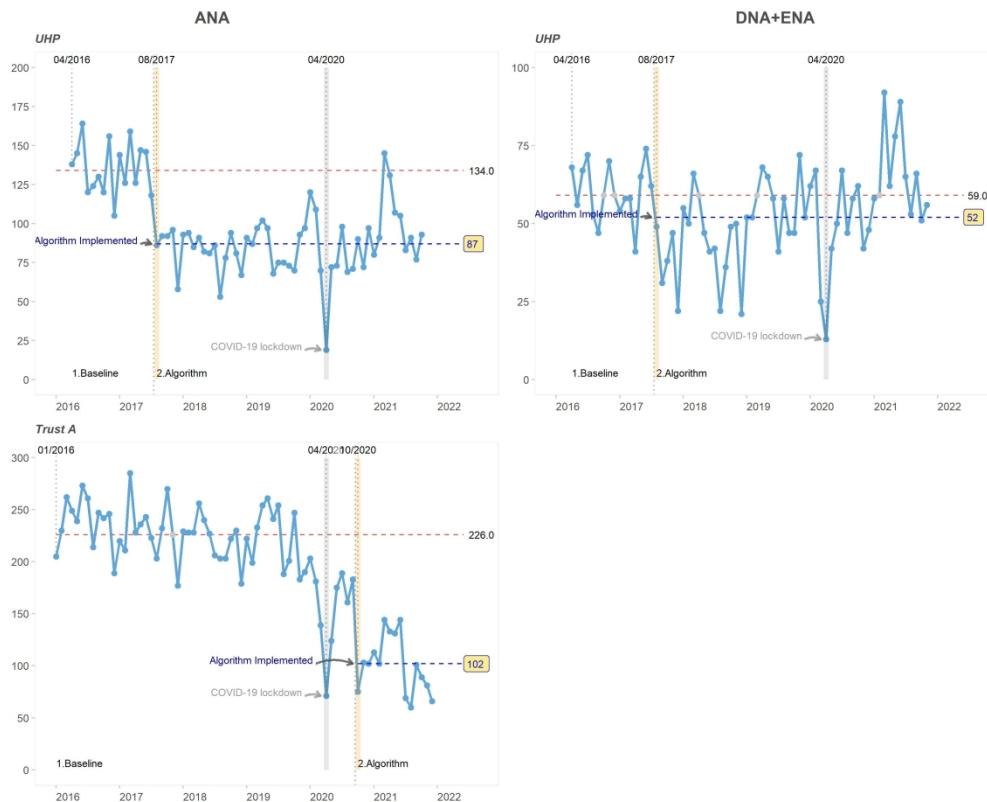
Figure 3: Run charts for monthly test activity for ANA (left) and DNA+ENA (right) tests at UHP (top row) and Trust B (data available only for ANA) The vertical gold bars are the first month when the algorithm for that test was in operation, separating prior (1.Baseline) from post (2.Algorithm) regimes. The baseline median activity is the red dotted line (projected forward for comparison; the post-implementation median demand is the blue dotted line (accompanied with the value in the gold-shaded box).

299x239mm (300 x 300 DPI)
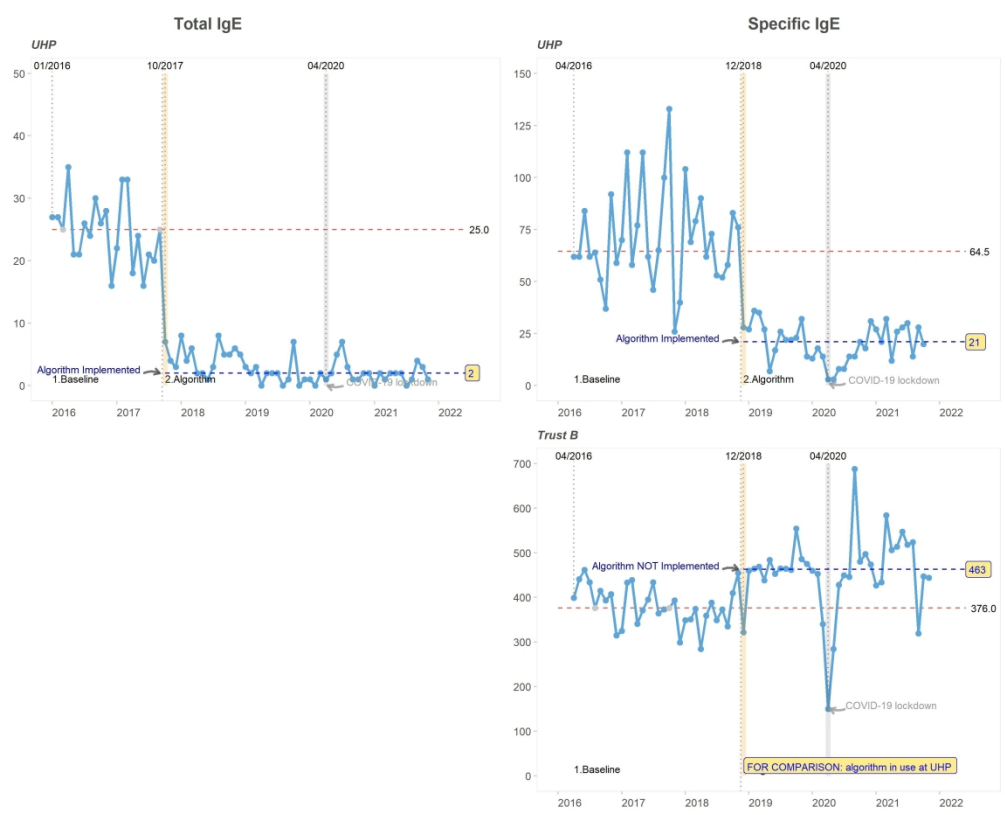
Figure 4: Run charts for monthly test activity for Total IgE and Allergen Specific IgE tests

299x239mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
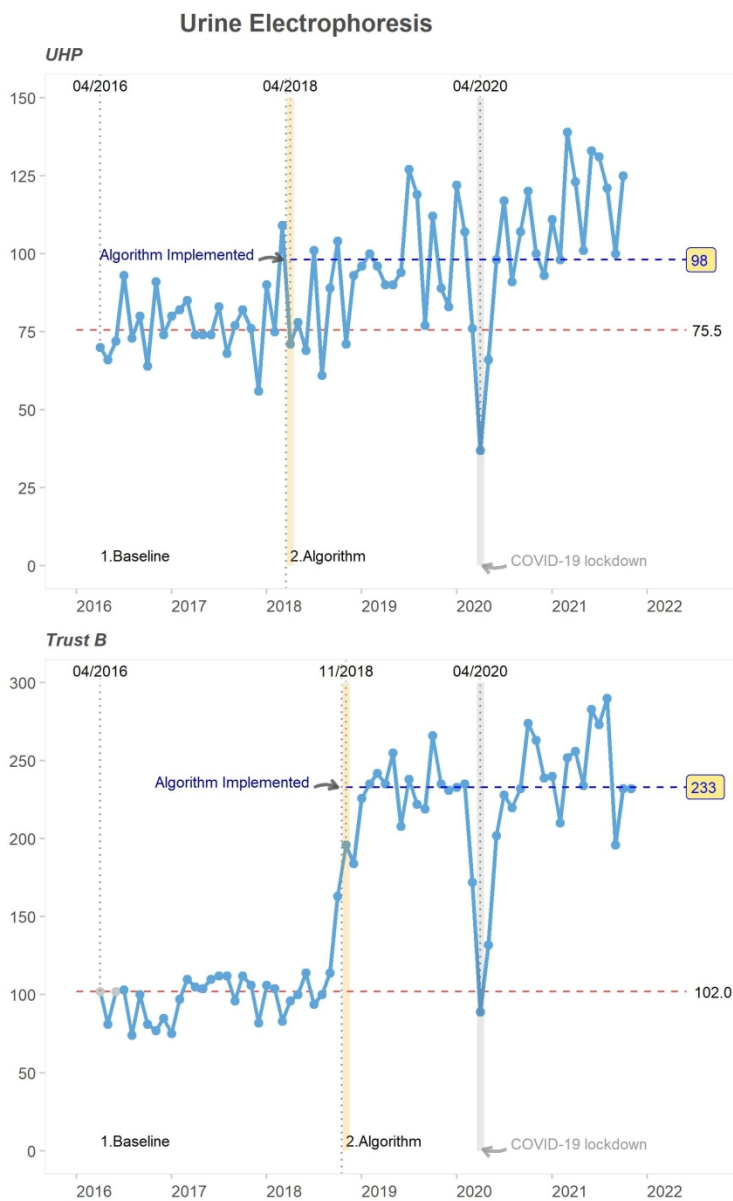
Figure 5: Run charts for monthly test activity for UE tests

149x239mm (300 x 300 DPI)

Figure 1: Process map of inappropriate test ordering process, showing potential demand management intervention points

Green = steps that are value to patient; Red = waste (non-value-adding or unnecessarily repeated steps); Orange line = necessary as a consequence of earlier waste

NOTE: for review we have added a 2nd supplementary pdf file containing copies of these main paper figures, which may be easier to read – Scholar one converts the 300dpi submission to low res when embedding them in the review document!

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

Figure 2: Driver Diagram showing hypothesised causal logic, options for intervention and points of measurement
(OM = outcome metric, PM = process metric, BM = balancing metric)

Figure 3 : Run charts for monthly test activity for ANA (left) and DNA+ENA (right) tests at UHP (top row) and Trust B (data available only for ANA)

The vertical gold bars are the first month when the algorithm for that test was in operation, separating prior (*1.Baseline*) from post (*2.Algorithm*) regimes. The baseline median activity is the red dotted line (projected forward for comparison; the post-implementation median demand is the blue dotted line (accompanied with the value in the gold-shaded box).

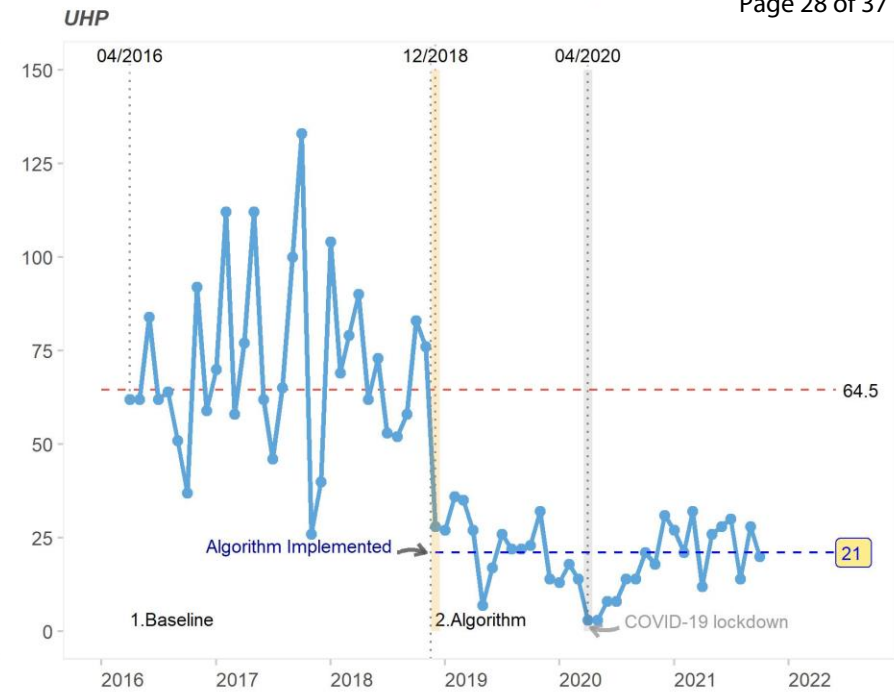**Figure 4 : Run charts for monthly test activity for Total IgE and Allergen Specific IgE tests**
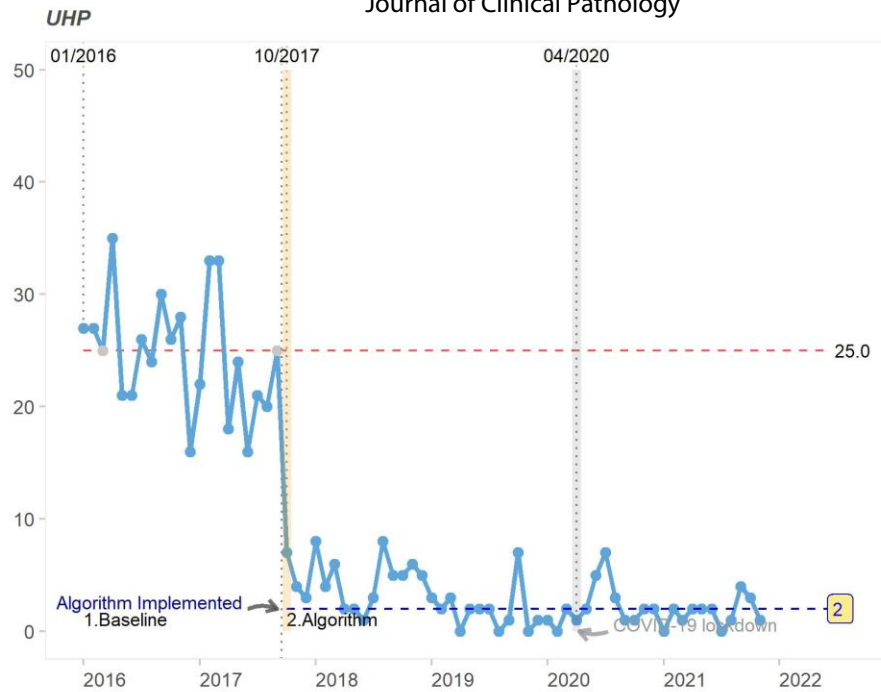
Figure 5 : Run charts for monthly test activity for UE tests

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
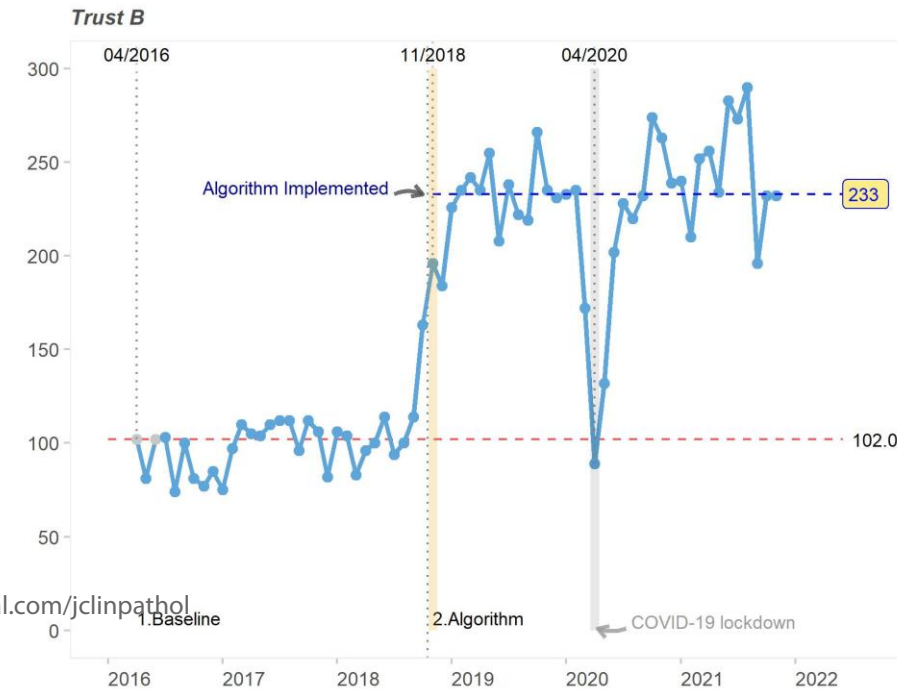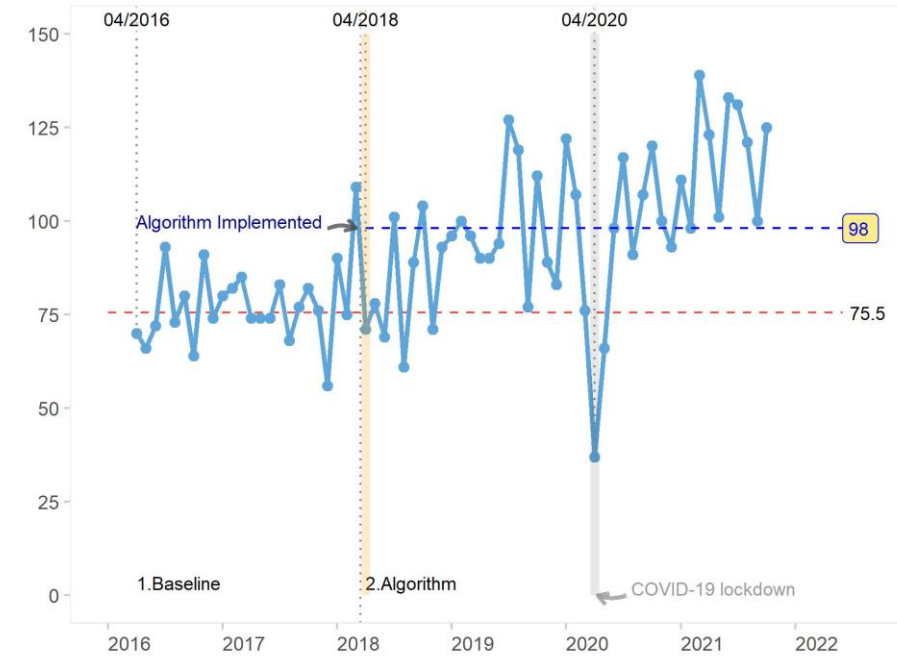27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

Urine Electrophoresis

SUPPLEMENTARY MATERIAL

APPENDIX 1 Operations Management

There has been a great emphasis in healthcare over recent decades to introduce quality improvement thinking and practice into healthcare. Much of this is based on the pioneering work of thinkers such as W. Edwards Deming, widely regarded as the father of quality management,[1] and on the iterative development of these ideas by companies such as Toyota, later generalised as 'lean'.[2]

In healthcare we can, as in for-profit sectors like automotive production, remove waste steps in the *flow* (e.g., in NHS labs[3-5] or operating theatres[6]) and waste of consumables such as valuable blood products.[7] However, there is often great potential to release capacity through removing waste from the *demand*, i.e., reducing unnecessary workload before it gets into our system and becomes activity, particularly in service sectors.[8]

In operations management terms, Waste C in Figure 1 of the main paper is termed 'failure demand': additional workload demand caused by the failure to do something correctly (or at all) for a 'customer' at a previous opportunity.[9] It has been investigated in many settings, including primary care and the police service.[10, 11].

In lean thinking, Wastes A and B are categorised as overproduction waste: producing output that is not required, at that moment or at all.[12]

The Intervention at Point B is 'quality inspection'. A key insight from Deming, which he famously demonstrated to many seminars of senior managers through his 'Red Bead Experiment' game, is not to try to inspect quality into a system: instead one must design quality in (and waste out) by understanding the root causes of waste and stopping them occurring in the first place.[13]

These are the central lessons that companies like Toyota took from Deming's seminars in Japan after WW2: cost and efficiency are in the *flow*, and non-value (defective or inappropriate) items must not be allowed to get into the flow to consume resources and generate even more work (e.g., rework and more failure demand). A well-known anecdote from the early investigations into Japanese production systems ('lean') was that it took Porsche more time to fix the defects found at final inspection than it took Toyota to build a perfect Lexus![14]

1

APPENDIX 2 GP Consultation: questionnaire survey

ANA questionnaire survey questions

1. How confident are you regarding when it is clinically appropriate to request an ANA test?
   a. Very confident
   b. Confident
   c. Unsure
   d. Very Unsure

2. How confident are you regarding the appropriate repeat interval for an ANA test?
   a. Very confident
   b. Confident
   c. Unsure
   d. Very Unsure

3. How often do you find info on ICE to be useful from a clinical perspective?
   a. Always
   b. Frequently
   c. Infrequently
   d. Never

4. How often do you consider the information on ICE to be useful for patients?
   a. Always
   b. Frequently
   c. Infrequently
   d. Never

We used the same questions pre- and post- implementation of the ANA requesting algorithm.

On both occasions we used Survey Monkey, with the link sent to 40 recent requestors of ANA tests. A total of 26 responses (40%) were received pre-implementation and 18 responses (28%) post-implementation.
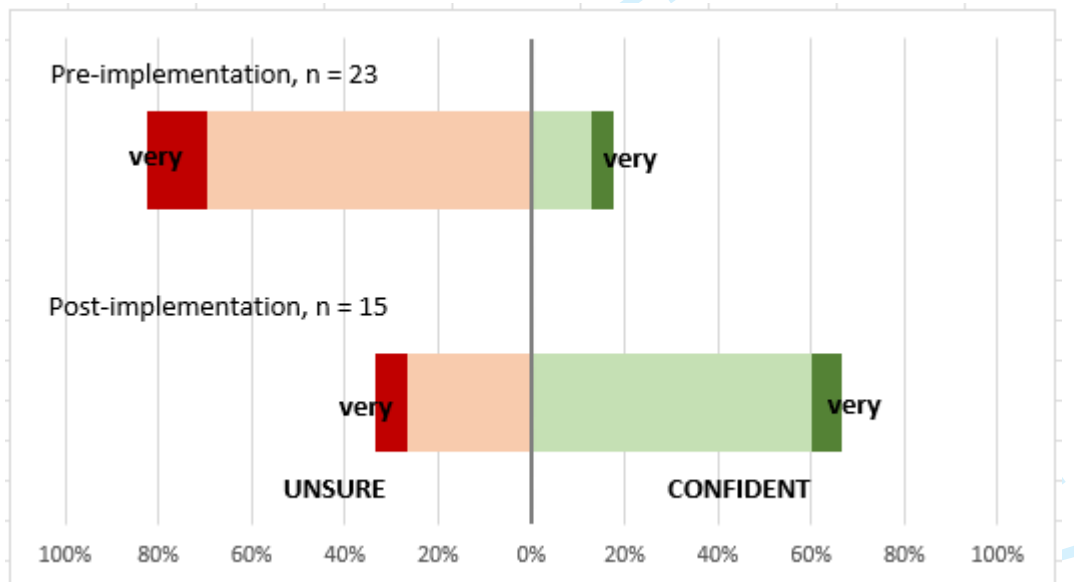
Responses

n = number of responses on that question. Total width of bar = 100%.
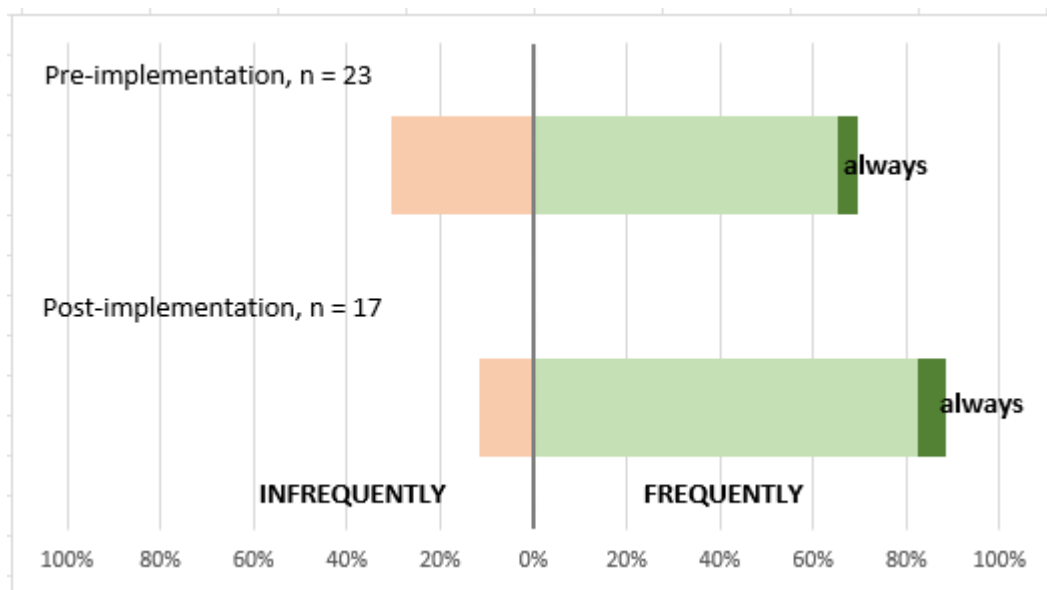
1. How confident are you regarding when it is clinically appropriate to request an ANA test?
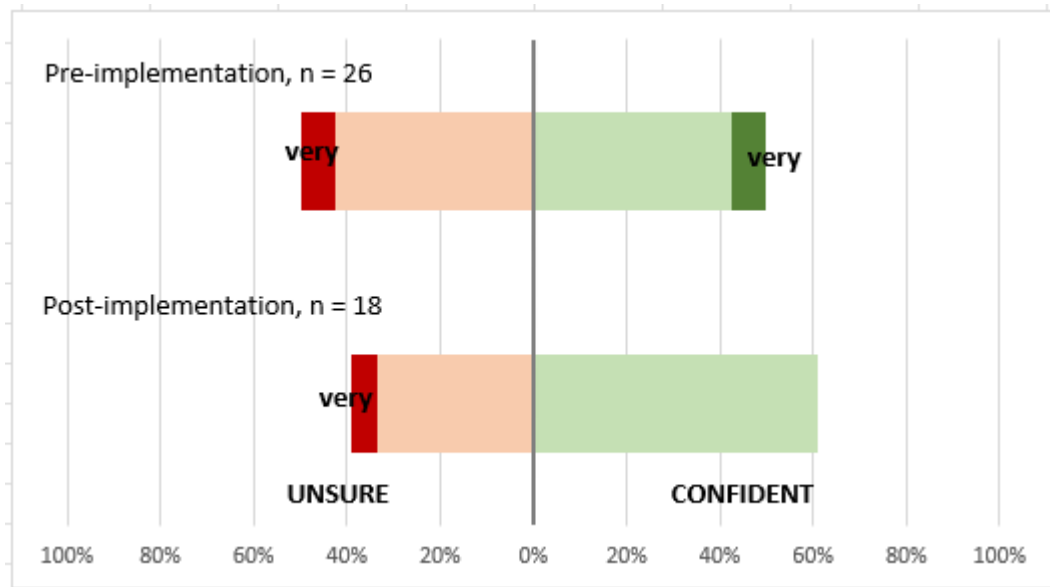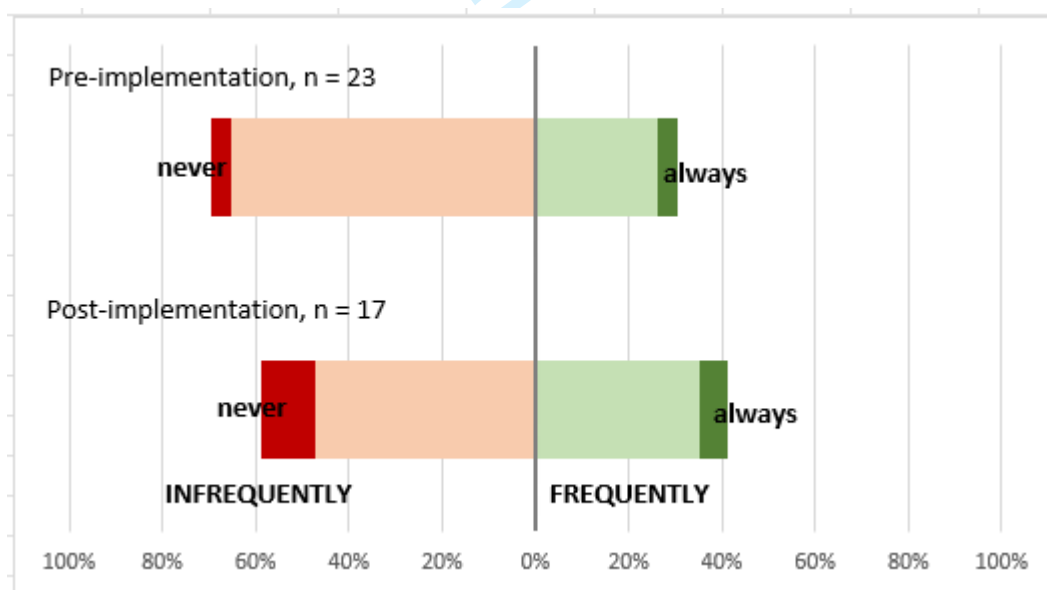

2. How confident are you regarding the appropriate repeat interval for an ANA test?

3

3. How often do you find info on ICE to be useful from a clinical perspective?

4. How often do you consider the information on ICE to be useful for patients?

Free-text responses were fragmented and tangential, with none taking the opportunity to express any dissatisfaction (other than the one comment noted in the main paper about ability to click on links – when they hadn't meant to!)

4

APPENDIX 3 Statistical Process Control (SPC) charts

Figures A3.1 and A3.2 are SPC Individuals (X) charts[15, 16] corresponding to the run charts in Figures 3-5.
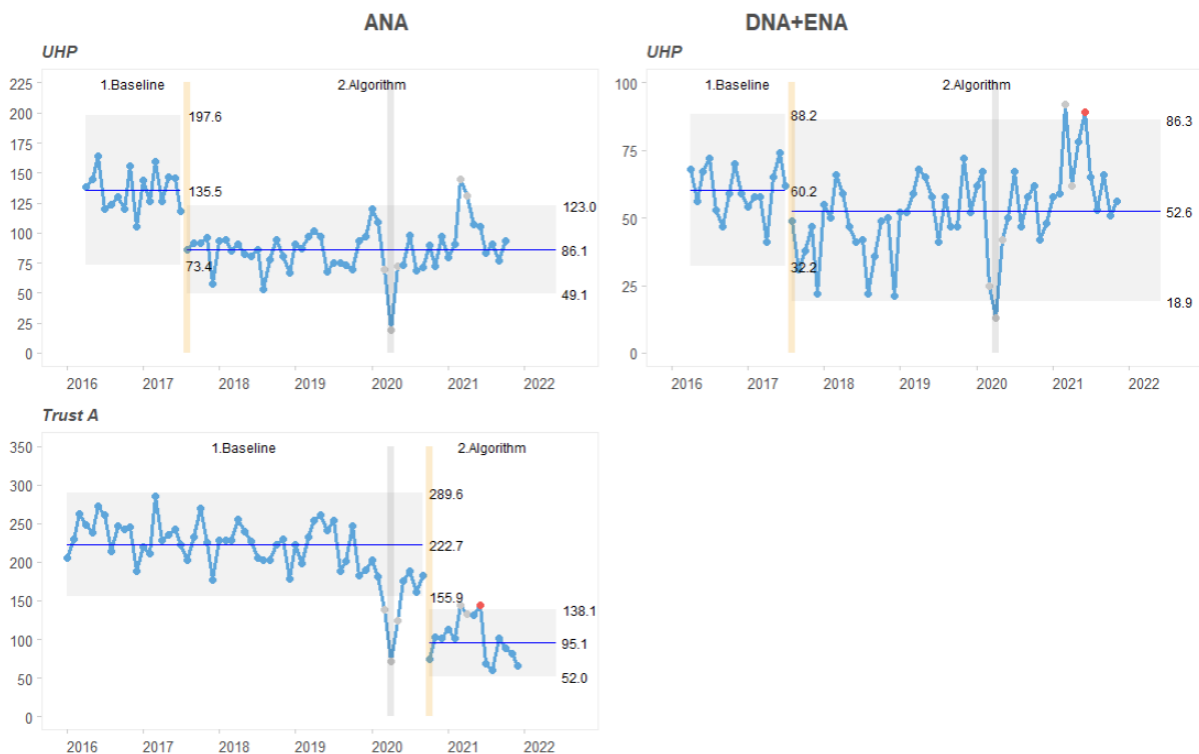

Figure A3.1 : SPC charts for monthly test activity for ANA and DNA+ENA tests

Note: the grey shaded horizontal ranges are not 95% confidence intervals on the *mean*, but are the 3 sigma limits (± 3 process standard deviations: the spread of the *individual* values). Sigma is estimated from a calculation involving the moving range.[15, 16]

The datapoints in grey are those excluded from the SPC and comparison of the means (including the t-tests). They are the periods of peak COVID-19 impact and an apparent 'catchup' in demand a year later. Though it is clear from the graphs that the COVID-19 impact was somewhat different for different tests and different trusts, for consistency we have applied this exclusion consistently to all datasets, excluding the five datapoints March, April & May 2020 and Match & April 2021. We have not applied this to the run charts and median, as these are more robust to extreme values. We can see from Table 2 that the percentage changes in these (trimmed) means are very similar to the percentage changes in the (untrimmed) medians. In SPC-terms, we exclude these points from analysis as 'special cause' variation (i.e, non-random or 'common cause' variation).

The datapoints in red are those (other than any grey ones) that exceed the ± 3 sigma process limits, i.e. are unusual from random variation. The underlying assumption of these charts are that the data are *roughly* normally distributed – the technique is robust to departures from this and it is not necessary to test normality. It is an empirical, robust, and visual technique. There are other rules for highlighting suspiciously non-normal behaviour, but this is the main one.[15-17] It is also worth noting that we have longer timeseries than is usual in SPC analysis, so rule violations ae more likely. The split at a known or potential regime change (e.g. intervention) recalculates the means and process limits, as shown.

We then observe that the process behaviour in our charts is consistent with random variation within each regime period. Then further, that there clearly *has* been a change in regime at the interventions. If we were to 'freeze' the mean and process limits at the baseline regime values and project them forward, we would see very many triggerings of the non-rando-behaviour-detection rules. The only exception in the graphs we present is the DNA+ENA activity above. We included this as a possible (beneficial) side-effect of reduced ANA activity. If we remove the split (knowledge that there has been an intervention), we would still see more triggering of the rules, but we can see visually that the change is much less decisive – not unexpectedly.
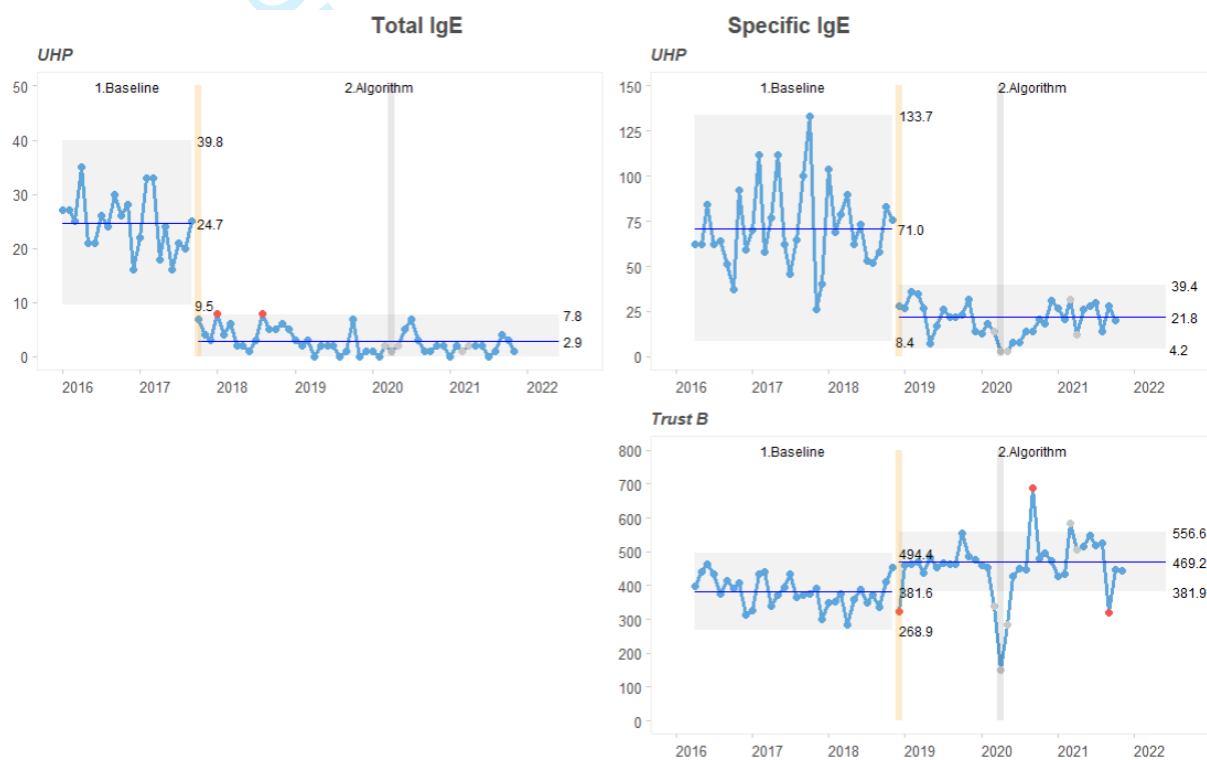

Figure A3.2 : SPC charts for monthly test activity for Total IgE and Specific IgE tests

Again, incontrovertible *reduction* in activity at UHP (top row) where the algorithms were implemented, not at Trust B. Activity at Trust B suggests an *increase*, but part of this may be a rapid bounce back from the first COVID-19 lockdown (spring 2020) and a spread-out catchup. Displacement of demand from UHP is not plausible, either logistically or contractually, and note that the size of the increase at Trust B is much larger than the reduction at UHP.
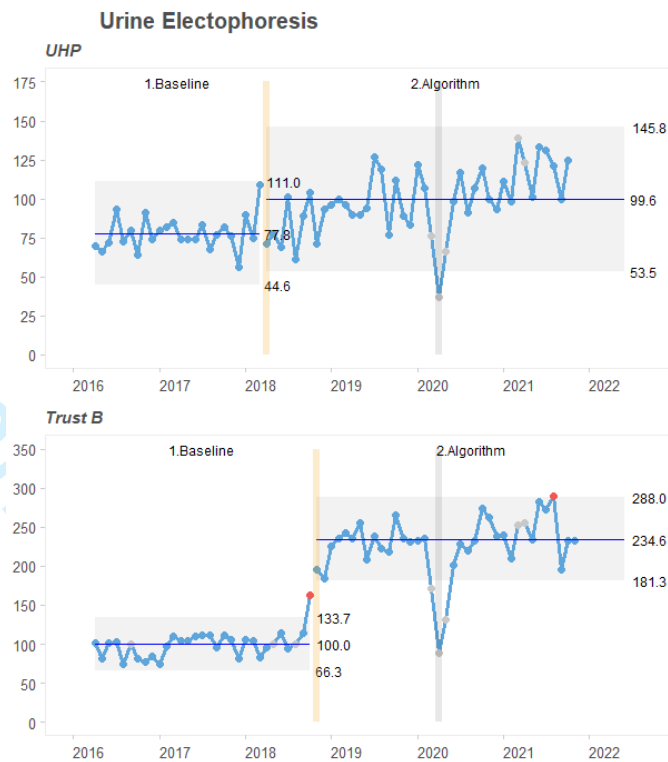
Figure A3.3 : SPC charts for monthly test activity for Urine Electrophoresis tests

Again, a clear shift of regime, particularly at Trust B. At UHP, the data suggests a gradual take-up, and possible continuing increase (though again this may be spread-out catchup from the COVID-19 lockdown.
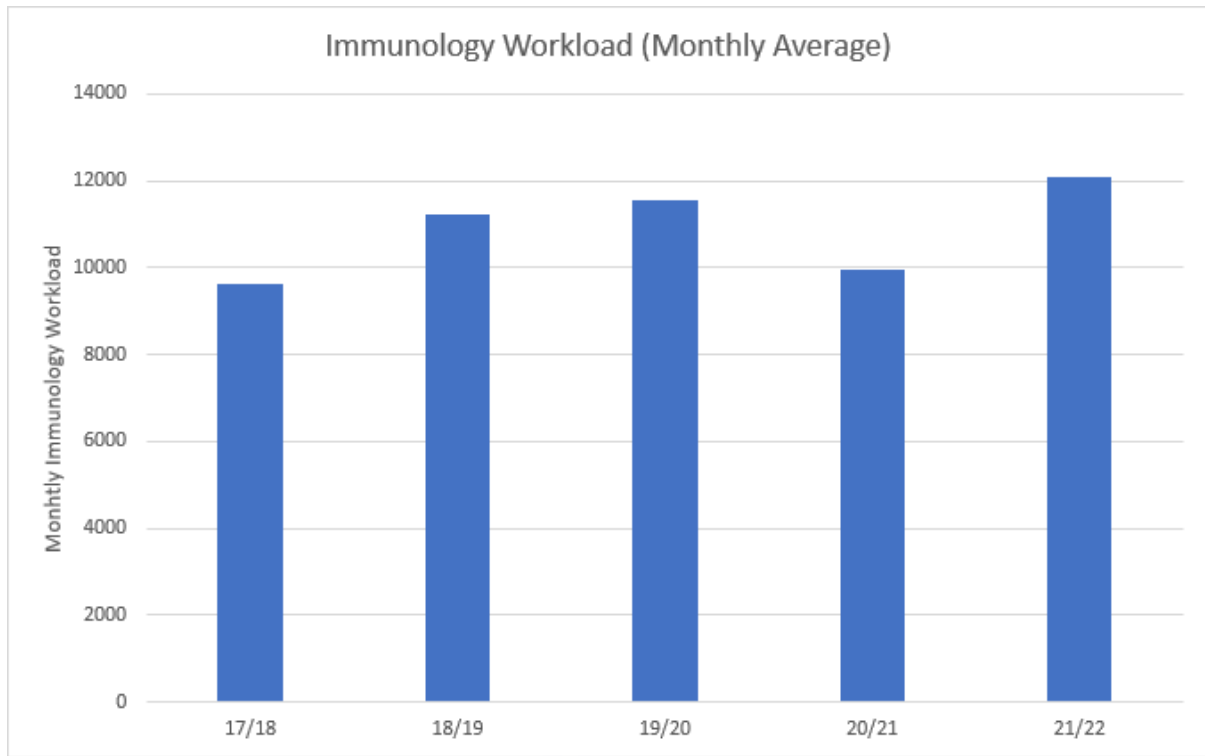
7

APPENDIX 4 – Balancing Metric

Figure A4.1 : UHP Immunology test workload (total tests conducted) – average monthly activity by financial year

References

1. Best M, Neuhauser D. W Edwards Deming: father of quality management, patient and composer. *Qual Saf Health Care* 2005;**14**(4):310-12. doi:10.1136/qshc.2005.015289
2. Womack JP, Jones DT. *Lean thinking: banish waste and create wealth in your corporation*. revised & updated ed. London: Simon and Schuster, 2003.
3. May F, Pepperall J, Davies E, et al. Summarised, verified and accessible: improving clinical information management for potential haematopoietic stem cell transplantation patients. *BMJ Open Qual* 2021;**10**(4):e001605. doi:10.1136/bmjoq-2021-001605
4. White E, Proudlove N, Kallon D. Improving turnaround times for HLA-B*27 and HLA-B*57:01 gene testing: A Barts Health NHS Trust Quality Improvement project. *BMJ Open Qual* 2021;**10**(3):e001538. doi:10.1136/bmjoq-2021-001538
5. Li Y, Proudlove N. A quality improvement project to reduce the turnaround times of infectious disease markers reporting in an NHS stem cell department. *BMJ Open Qual* 2022;**11**(2):e001814. doi:10.1136/bmjoq-2022-001814
6. Pridgeon M, Proudlove N. Getting going on time: reducing neurophysiology setup times in order to contribute to improving surgery start and finish times. *BMJ Open Qual* 2022;**11**(3):e001808. doi:10.1136/bmjoq-2021-001808
7. McCullagh J, Proudlove N, Tucker H, et al. Making every drop count: reducing wastage of a novel blood component for transfusion of trauma patients. *BMJ Open Qual* 2021;**10**(3):e001396. doi:10.1136/bmjoq-2021-001396
8. Seddon J. *Freedom from command and control: a better way to make the work work... the Toyota system for service organisations*. 2nd ed. Buckingham, UK: Vanguard Education, 2005.
9. Seddon J. *Systems thinking in the public sector*. Axminster, UK: Triarchy Press, 2008.
10. Walley P, Found P, Williams S. Failure demand: a concept evaluation in UK primary care. *Int J Health Care Qual Assur* 2019;**32**(1):21-33. doi:10.1108/IJHCQA-08-2017-0159
11. Morris G, Walley P. Implementing failure demand reduction as part of a demand management strategy. *Public Money Manag* 2022;**42**(1):22-31. doi:10.1080/09540962.2021.1978163
12. Boaden R, Harvey G, Moxham C, et al. Quality improvement: theory and practice in healthcare. 2008. Coventry: NHS Institute for Innovation and Improvement. www.england.nhs.uk/improvement-hub/wp-content/uploads/sites/44/2017/11/Quality-Improvement-Theory-and-Practice-in-Healthcare.pdf. (accessed 12 August 2021).
13. Deming WE. *Out of the crisis*. Cambridge, Mass.: Centre of Advanced Engineering Study, MIT, 1986.
14. Womack JP, Jones DT, Roos D. *The machine that changed the world*. New York: Rawson Associates, 1990.
15. Provost LP, Murray SK. *The health care data guide: learning for data improvement*. San Francisco, CA: Jossey-Bass, 2011.
16. Mohammed MA, Worthington P, Woodall W. Plotting basic control charts: tutorial notes for healthcare practitioners. *Qual Saf Health Care* 2008;**17**(2):137-45. doi:10.1136/qshc.2004.012047
17. Anhøj J, Wentzel-Larsen T. Sense and sensibility: on the diagnostic value of control chart rules for detection of shifts in time series data. *BMC Med Res Methodol* 2018;**18**(1):100. doi:10.1186/s12874-018-0564-0