BMJ open
accessible medical research

What is the best strategy for investigating abnormal liver function tests in primary care? Implications from a prospective study

Richard J Lilford,[1] Louise M Bentham,[1] Matthew J Armstrong,[2] James Neuberger,[3] Alan J Girling[1]

To cite: Lilford RJ, Bentham LM, Armstrong MJ, *et al*. What is the best strategy for investigating abnormal liver function tests in primary care? Implications from a prospective study. *BMJ Open* 2013;**3**:e003099. doi:10.1136/bmjopen-2013-003099

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-003099).

Received 22 April 2013
Accepted 13 May 2013

[1]School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham, UK
[2]Centre for Liver Research and NIHR Biomedical Research Unit, University of Birmingham, Edgbaston, Birmingham, UK
[3]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Correspondence to
Professor Richard J Lilford; r.j.lilford@bham.ac.uk

ABSTRACT

Objective: Evaluation of predictive value of liver function tests (LFTs) for the detection of liver-related disease in primary care.

Design: A prospective observational study.

Setting: 11 UK primary care practices.

Participants: Patients (n=1290) with an abnormal eight-panel LFT (but no previously diagnosed liver disease).

Main outcome measures: Patients were investigated by recording clinical features, and repeating LFTs, specific tests for individual liver diseases, and abdominal ultrasound scan. Patients were characterised as having: hepatocellular disease; biliary disease; tumours of the hepato-biliary system and none of the above. The relationship between LFT results and disease categories was evaluated by stepwise regression and logistic discrimination, with adjustment for demographic and clinical factors. True and False Positives generated by all possible LFT combinations were compared with a view towards optimising the choice of analytes in the routine LFT panel.

Results: Regression methods showed that alanine aminotransferase (ALT) was associated with hepatocellular disease (32 patients), while alkaline phosphatase (ALP) was associated with biliary disease (12 patients) and tumours of the hepatobiliary system (9 patients). A restricted panel of ALT and ALP was an efficient choice of analytes, comparing favourably with the complete panel of eight analytes, provided that 48 False Positives can be tolerated to obtain one additional True Positive. Repeating a complete panel in response to an abnormal reading is not the optimal strategy.

Conclusions: The LFT panel can be restricted to ALT and ALP when the purpose of testing is to exclude liver disease in primary care.

ARTICLE SUMMARY

Article focus

- The response to an abnormal liver function test (LFT) result in primary care is highly eclectic.
- Guidelines suggest repeating an abnormal standard LFT panel.
- We conducted a prospective study to evaluate the prognostic value of LFTs.

Key messages

- The prevalence of significant liver disease in people with incidental abnormal LFTs is little higher than the population prevalence.
- The policy of requesting a standard LFT panel with a view to repeating it if abnormal is inefficient.
- Just two analytes (alanine aminotransferase and alkaline phosphatase) provide an efficient default testing strategy for excluding liver disease (of viral, genetic, autoimmune or neoplastic origin) in primary care.

Strengths and limitations of this study

- This is the first large, prospective primary care-based study of patients with abnormal LFTs that were fully evaluated for liver disease.
- Patients where all analytes were normal were not included, meaning that while True and False-Positive rates are unbiased, sensitivity and specificity may be overestimated and underestimated, respectively.

INTRODUCTION

Liver function tests (LFTs) are inexpensive tests that are frequently ordered in a panel of up to eight analytes as a 'test of exclusion' in patients with non-specific symptoms or as part of routine health checks. LFTs are difficult to study because the tests portend a very large number of diseases, some of them very rare. Nevertheless, there is a large literature on LFTs; a review by Green and Flamm located 6000 papers published since 1990 alone.[1] However, this literature mostly originates from hospital practice and often deals with a restricted number of analytes. Moreover, it is predominantly retrospective and concerned with the probabilities of test results given the disease-state, whereas the

clinician typically starts with the LFT result and needs to know the predictive probability of the disease. An updated review of the literature[2] shows that there are no prospective studies based in primary care practice where patients were fully investigated following at least one abnormal analyte from a full LFT panel. It is therefore not surprising that eclectic decision-making has been documented in primary care.[3] Birmingham and Lambeth Liver Evaluation Testing Strategies (BALLETS) was a prospective UK study that aimed to assess the value of abnormal LFT analytes for predicting significant liver disease in primary care. The detailed report of the study will appear in the Health Technology Assessment (HTA) monograph series.[2] Here, we use the study information to investigate the diagnostic potential of LFT results, taking account of the individual patient characteristics, and examine the positive predictive performance of different LFT panels for the diagnosis of liver disease using standard laboratory-based reference ranges that inform general practitioner (GP) decision-making. We consider viral, genetic and autoimmune diseases and tumours of the hepatobiliary system; a discussion of fatty liver will appear in the HTA report.[2]

METHODS

BALLETS study

Data collection

BALLETS was a prospective UK study of patients with an abnormal LFT panel across eight primary care practices in Birmingham and three in the Lambeth area of London. The 11 practices were served by three laboratories following similar analytical procedures, one of which accounted for over 80% of the sample.[2] Patients were eligible for the study if they did not have obvious or pre-existing liver disease, and one or more of the eight analytes in an index LFT panel was abnormal. We set out to recruit 1500 such patients on the grounds that this would allow us to examine the predictive performance for liver disease of up to 12 variables without overfitting using a 10 to 1 'events per variable' rule. This calculation was supported by computing the chance of missing high-risk cases when using a logistic discriminant function based on LFTs.[2] The index panel comprised: alanine aminotransferase (ALT), aspartate aminotransferase (AST), γ-glutamyltransferase (GGT), bilirubin (Bili), alkaline phosphatase (ALP), albumin (Alb), globulin (Glob), and total protein (Tprot).[4] Analyte abnormality was determined using standard laboratory reference ranges, which are routinely adjusted for age and gender where appropriate.[2]

Recruitment took place from 2005 to 2008. Eligible patients were invited to join the study and attend a first follow-up session (FU1) at the practice where the following data were collected:

1. Patient and clinical characteristics, including age, sex, ethnic group, country of birth, reason for blood testing, medication and history of illness, substance abuse, travel, immunisation and transfusion history;
2. Alcohol use, via a standardised questionnaire[2];
3. Weight, height, waist and hip circumference measurements;
4. Repeat of the eight-analyte LFT panel;
5. Blood for specific (auto-immune, genetic and viral) diseases in the 'liver work up' (table 1);
6. Ultrasound scan (USS) of the upper abdomen. Any tumours of the hepatobiliary system were noted and the liver was classified as normal, echobright (in three levels of intensity) or cirrhotic. A sample of ultrasound films were reviewed by the study radiologist.

The research team produced a consolidated report comprising the results of the index LFT panel and the information collected from follow-up. The patient then attended the primary care practitioner for a consultation informed by the consolidated report. Participating primary care practitioners were provided with a set of guidelines[2] to assist in future decision-making when one of the tests in table 1 was abnormal, or when an abnormality was seen on USS. Primary care practice and hospital records were reviewed by the research team to harvest information gleaned from follow-up tests. Primary care practitioners were alerted if follow-up investigations had not been carried out when indicated.

Diagnostic categories

The number of diseases that might cause abnormal LFT results is very large. This issue was tackled by grouping the diseases into categories that made sense clinically and pathophysiologically (table 1). These were:

1. Hepatocellular disease
2. Biliary disease
3. Tumours of the hepatobiliary system

All other patients were placed in a 'non-specific' category.

Follow-up

The BALLETS cohort was followed up by examination of the primary care and hospital records and a follow-up visit after 2 years (FU2), where clinical examination (repeat LFTs and abdominal ultrasound) was repeated. The results of the 2-year follow-up are included in the full report.[2]

Analytical approach

Exploratory analysis of analyte concentrations

A hierarchical stepwise approach was used to investigate between-patient variation in each of the eight analyte concentrations in the FU1 panel, log-transformed to improve distributional symmetry. All analyses were adjusted for laboratory effects. First, the log-concentration was described by a linear (analysis of variance) model using the main effects of age, sex, ethnic group, body mass index (BMI) and alcohol

Table 1 Liver disease (viral, genetic and autoimmune) for which all patients were tested*

Category	Disease	Blood tests carried out on all members of the cohort (to diagnose or screen for the disease)	Method by which diagnosis was made in screen positive cases
Hepatocellular diseases	Chronic viral hepatitis C	Hepatitis C virus antibody (HCV Ab)	Viral marker positive and hepatologist's opinion
	Chronic viral hepatitis B	Hepatitis B surface viral antigen (HBV surface Ag)	Viral marker positive and hepatologist's opinion
	Metal storage disease: Iron (haemochromatosis)	Transferrin levels	Genotype performed on patient if transferrin saturation >50%
	Autoimmune hepatitis	Smooth muscle antibody	Raised antibodies and ALT or AST or globulin exceeding twice the upper limit of normal. Confirmed by the hepatologist's opinion
	Metal storage disease: Copper (Wilson's disease)	Caeruloplasmin	Low levels of caeruloplasmin and hepatologist's opinion
	α-1 Antitrypsin (A1AT) deficiency	A1AT level	Phenotype testing performed if A1AT abnormal
	Alcoholic/fat-induced cirrhosis or hepatocellular cancer (HCC)	N/A	Abdominal ultrasound+exclusion of other diseases in this table and hepatologist's opinion
Intrahepatobiliary duct disease	Primary biliary cirrhosis (PBC)	Antimitochondrial antibody	Anti-mitochondrial antibodies (≥1:40 titre) and hepatologist's opinion
	Primary sclerosing cholangitis	N/A	Combination raised ALP and ulcerative colitis. Confirmed by hepatologist's opinion

Previously undiagnosed cirrhosis of other causes was also included in this category.
*We did not include the benign condition Gilbert's syndrome in any disease category.

consumption (table 2) together with all two-way interactions involving age or sex. Backwards elimination was applied first to remove non-significant ($p>0.05$) interactions and then to remove non-significant main effects from the model, with the provisos that the main effects of age and sex, and the age-by-sex interaction were always retained, and that no main effect was removed if it featured in a significant interaction. The threshold for

Table 2 Baseline patient characteristics (N=1290)

Reason for testing	Signs and Symptoms 406 (31.5)	Chronic disease review 884 (68.5)				
Sex	Male 724 (56.1)	Female 566 (43.9)				
Age (years)	≤34 106 (8.2)	35–44 165 (12.8)	45–54 240 (18.6)	55–64 325 (25.2)	65–74 273 (21.2)	75+ 181 (14.0)	
Ethnic group	White 1056 (81.9)	Asian 89 (6.9)	Black 66 (5.1)	Other 40 (3.1)			Not known 39 (3.0)
Country of birth	UK 1022 (79.2)	Indian Subcontinent 60 (4.7)	Other countries 180 (14.0)				Not known 28 (2.2)
BMI at FU1 (kg/m^2)	<20 49 (3.8)	20–24.99 250 (19.4)	25–29.99 454 (35.2)	≥30 498 (38.6)			Not known 39 (3.0)
Alcohol at FU1 (units/week*)	0 547 (42.4)	1–14 352 (27.8)	15–29 153 (11.9)	30–49 122 (9.5)	50–99 84 (6.5)	100+ 24 (1.9)	Not known 8 (0.6)

Entries are frequencies (and per cent of total).
*1 unit=10 g of alcohol, FU1: follow-up visit 1.
BMI, body mass index.

exclusion was set relatively high ($p>0.05$), which tends to increase the explained variation in LFTs, thus reducing the risk of finding marginal differences between diagnostic groupings that could be attributed to patient characteristics. These analyses were confined to patients for whom a complete set of patient characteristics had been recorded (N=1211). Finally, the marginal impact of the diagnostic category was investigated by adding factors representing the five diagnoses of: viral hepatitis; other hepatocellular disease; biliary disease; hepato-biliary tumour and non-specific.

Diagnostic potential

Analyte concentrations were first scaled to the laboratory that performed the largest proportion of tests (78%), using factors estimated in the exploratory analysis. Stepwise logistic regression was used to determine the best combinations of patient characteristics and (scaled) analyte concentrations to distinguish between the non-specific diagnostic group and each of the three main liver disease groups in separate analyses. The analysis was repeated for a subcategory of hepatocellular disease, viral hepatitis, because of its clinical importance. The candidate variables were: age, sex, ethnic group, BMI, country of birth and all eight analyte concentrations (logged) from the FU1 follow-up panel. Interactions were not considered. Missing values in all candidate variables were handled using the chained equation method in Stata V.12.[5] Significant predictors were identified using four complementary procedures: backward elimination, with a $p>0.01$ threshold for exclusion from the model; forward selection, with $p<0.01$ for inclusion and two mixed forward and backward procedures with $p>0.01$ for exclusion and $p<0.005$ for inclusion.

Comparison of index panels for liver disease diagnosis

The absence of patients with completely normal analyte concentrations in the index sample (as determined by conventional reference ranges) precludes a comprehensive analysis of the diagnostic performance of LFTs. Positive predictive performance was addressed using the laboratory-based reference ranges commonly used in general practice, though it is impossible to consider the impact of relaxing the thresholds for abnormality in this data set. The analysis considers the 255 ($=2^8-1$) possible index LFT panels that can be constructed from eight analytes and is confined to the 915 patients with index measurements on all eight analytes. For a particular patient, a panel was considered to be positive if at least one analyte concentration fell outside its reference range. A positive panel was characterised as a True Positive (TP) or False Positive (FP) according to whether it belonged to a patient with or without liver-related disease, defined broadly to include all serious diseases (hepatocellular, biliary and tumours of the hepatobiliary system (categories 1, 2 and 3)). One panel dominates another if it generates more TPs and fewer FPs. Otherwise, preference between two panels can be determined if the trade-off between the value of a TP and the cost of an FP is specified. A panel with more TPs and FPs than another panel will be preferred if the ratio of the extra TPs to FPs generated is more than the trade-off value. For example, if the TP/FP trade-off is 0.01, then finding one extra TP can compensate for incorrectly identifying up to 100 extra FPs. SEs for the ratios of extra TPs and FPs for comparing pairs of panels, and for the Positive Predictive Values (PPVs) of individual panels, were obtained from 1000 bootstrap resamples from the original 915 patients.

Treatment of missing data

The analysis of alternative index panels is restricted to patients with a complete set of eight index analytes. Including incomplete panels here would mean that the yield (total number of positives) from a subpanel of fewer than eight analytes would be biased upwards, since some patients would owe their presence in the study to abnormalities on just those analytes (and no others). Nevertheless, restriction to complete index panels cannot eliminate all recruitment biases since it favours those practices where the GPs have applied the study protocol most attentively. The analytical choices have been made, to some extent, on pragmatic grounds. Thus, the exploratory regression and discriminant analyses of the FU1 panel have been applied to all patients in order to maximise coverage. In any case, the potential for bias is reduced here because of the imperfect correlation between index and follow-up tests, and because the analysis does not refer explicitly to the thresholds of abnormality that triggered recruitment to the study. Incompleteness in FU1 panels is uninformative since it results from the laboratory's failure to report rather than the GP or patient's non-compliance.

RESULTS

Patients and data

The study sample of 1290 patients is summarised in table 2. Index panels were available for all 1290 patients, of which 915 (70.9%) included all eight analytes. The FU1 panel was taken after a median of 30 days postindex (IQR 21–51). There were 1275 patients (98.8%) with an LFT panel at follow-up, of which 1168 (92%) were complete. Eighty-five per cent (992/1168) of complete FU1 panels had an abnormal LFT, falling slightly to 84% (706/844) where the index panel was also complete. The correlation between index and follow-up tests was high for all analytes, ranging from 0.66 for Tprot to 0.89 for GGT. Hence, the initial level of abnormality has a marked influence on the probability that an abnormal analyte will revert to normal on repeat testing. The five non-protein analytes (ALT, AST, Bili, ALP, GGT) all showed a reduction over time that might be interpreted as a regression to the mean (table 3).

Table 3 Analyte concentrations and abnormalities by diagnostic category

					Abnormalities by diagnostic category: number abnormal/number tested (%)				
Analyte (units)	**Panel**	**N**	**Median**	**IQR**	**Total (N=1290)**	**Non-specific (N=1237)**	**Category 1 (N=32)**	**Category 2 (N=12)**	**Category 3 (N=9)**
ALT	Index	1114	34	(22–52)	438/1114 (39.3)	415/1071 (38.8)	18 /27 (66.7)	3/8 (37.5)	2/8 (25.0)
(U/l)	FU1	1234	31	(22–46)	375/1234 (30.4)	346/1184 (29.2)	23/30 (76.7)	3/11 (27.3)	3/9 (33.3)
AST	Index	1158	29	(23–40)	255/1158 (22.0)	237/1108 (21.4)	14/29 (48.3)	3/12 (25.0)	1/9 (11.1)
(U/l)	FU1	1212	28	(23–37)	172/1212 (14.2)	153/1163 (13.2)	15/30 (50.0)	3/11 (27.3)	1/8 (12.5)
Bili	Index	1265	9	(7–13)	148/1265 (11.7)	142/1213 (11.7)	5/31 (16.1)	1/12 (8.3)	0/9 (0.0)
(μmol/L)	FU1	1233	9	(6–13)	111/1233 (9.0)	106/1185 (9.0)	3/29 (10.3)	1/11 (9.1)	1/8 (12.5)
ALP	Index	1272	188	(144–247)	189/1272 (14.9)	172/1220 (14.1)	5/31 (16.1)	9/12 (75.0)	3/9 (33.3)
(U/l)	FU1	1236	187	(142–238)	143/1236 (11.6)	130/1188 (10.9)	4/29 (13.8)	7/11 (63.6)	2/8 (25.0)
GGT	Index	1152	64.5	(44–104)	867/1152 (75.3)	833/1108 (75.2)	18/28 (64.3)	8/8 (100.0)	8/8 (100.0)
(U/l)	FU1	1243	58	(37–98)	787/1243 (63.3)	749/1193 (62.8)	20/31 (64.5)	9/10 (90.0)	9/9 (100.0)
Alb	Index	1278	45	(43–47)	30/1278 (2.4)	29/1225 (2.4)	1/32 (3.1)	0/12 (0.0)	0/9 (0.0)
(g/l)	FU1	1254	46	(44–48)	40/1254 (3.2)	36/1206 (3.0)	4/29 (13.8)	0/11 (0.0)	0/8 (0.0)
Glob	Index	977	29	(27–32)	55/977 (5.6)	53/938 (5.7)	2/23 (8.7)	0/8 (0.0)	0/8 (0.0)
(g/l)	FU1	1214	30	(27–33)	74/1214 (6.1)	66/1167 (5.7)	4/28 (14.3)	3/11 (27.3)	1/8 (12.5)
Tprot	Index	981	74	(71–77)	97/981 (9.9)	93/942 (9.9)	4/23 (17.4)	0/8 (0.0)	0/8 (0.0)
(g/l)	FU1	1235	76	(73–79)	199/1235 (16.1)	187/1185 (15.8)	9/30 (30.0)	2/11 (18.2)	1/9 (11.1)

FU1: follow-up visit 1 (mean of 30-day postindex bloods).
ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; Bili, bilirubin; GGT, γ-glutamyltransferase; Glob, globulin; Tprot, total protein.

Prevalence of disease in the cohort

Hepatocellular diseases were present in 32 cases (2.5%) and biliary disease in 12 cases (0.9%). Viral hepatitis B or C was the most common hepatocellular disease (13 cases, all subsequently treated), followed by haemochromatosis (four compound heterozygote and six homozygous, of whom four were treated by regular venesection), cirrhosis (six cases, including one case of hepatoceullular carcinoma) and α 1-antitrypsin deficiency (three cases). Biliary diseases comprised Primary Biliary Cirrhosis (10 cases) and Primary Sclerosing Cholangitis (two cases). Tumours of the hepatobiliary system (nine cases) were metastatic liver cancer (four cases), cancer of the pancreas or bile duct (four cases) and amoebic liver abscess (one case).

Analyte concentrations, patient characteristics and disease category

The results of the stepwise regressions are summarised in table 4. The main effects and interaction for Age and Sex were included, by design, in all base models. The relationship with BMI was significant for all analytes except ALP, and varied with age except in the case of Bili. The effect of alcohol was significant for ALT, AST and GGT. The ethnic group impacted on protein analytes most markedly on Glob levels, which were raised in non-white groups (see the main report for fuller details).[2] The impact of disease categories is presented in terms of multiplicative factors applied to the analyte concentrations (table 4). Significant effects are evident for ALT and AST (both raised in hepatocellular disease); ALP (raised in biliary disease and tumours of the hepatobiliary system); GGT and Glob (raised in biliary disease) and Alb (reduced in tumours).

Diagnostic potential

For both the biliary disease and hepatobiliary tumour categories, all four stepwise procedures converged to the same single diagnostic indicator, namely ALP, with no other analytes or patient characteristics retained in the models. The associated c-statistics were 0.84 for biliary disease and 0.83 for hepatobiliary tumours. For the diagnosis of hepatocellular diseases, ALT and AST emerged as alternative diagnostic markers, depending on the details of the stepwise procedure. The alternative models were ALT with BMI (c-statistic 0.80) and AST with Country of Birth (c-statistic 0.76). When the viral hepatitis subgroup was considered as a separate category, Country of Birth featured in all models, alongside one or the other of these two analytes with similar c-statistics 0.92 (ALT with Country of Birth) and 0.89 (AST with Country of Birth). A further analysis was performed, contrasting non-hepatitis hepatocellular disease with the non-specific group. Here, AST and ALT again emerged as alternatives, with near identical c-statistics (=0.76 to 2 dp), but no other variables were retained by any of the four stepwise procedures. Thus, only three analytes featured in the individual diagnostic models: ALT, AST and ALP.

Table 4 Summary of exploratory of regression models

Analyte	Covariates in base model (agexsex included)	Percentage of variance explained (base model)	Impact of disease category (multiplicative factors and 95% CI)			Number of cases used (% of study sample)
			Disease category 1		Disease category 2	Disease category 3	
			Hepatitis B or C	Other			
ALT	BMIxage; alcohol	19.7	2.25 (1.65 to 3.07)	1.61 (1.27 to 2.04)	1.07 (0.79 to 1.46)	0.98 (0.70 to 1.37)	1159 (89.8)
AST	BMIxage; alcohol	6.7	1.69 (1.35 to 2.13)	1.56 (1.30 to 1.88)	1.20 (0.94 to 1.51)	0.93 (0.70 to 1.22)	1138 (88.2)
Bili	BMI; alcohol	9.8	0.93 (0.68 to 1.28)	1.27 (0.99 to 1.64)	0.83 (0.60 to 1.14)	0.92 (0.64 to 1.33)	1156 (89.6)
ALP	Alcohol	9.3	0.94 (0.77 to 1.15)	0.98 (0.84 to 1.15)	1.64 (1.35 to 1.98)	1.54 (1.22 to 1.95)	1159 (89.8)
GGT	BMIxage; alcohol	13.8	1.17 (0.76 to 1.79)	1.51 (1.08 to 2.13)	1.68 (1.06 to 2.68)	1.22 (0.75 to 2.00)	1167 (90.5)
Alb	BMIxage; ethnicity	11.3	1.02 (0.97 to 1.06)	1.00 (0.97 to 1.04)	0.97 (0.93 to 1.01)	0.93 (0.89 to 0.98)	1176 (91.2)
Glob	BMIxage; ethnicity	8.7	1.05 (0.95 to 1.15)	0.97 (0.90 to 1.05)	1.13 (1.03 to 1.24)	1.02 (0.91 to 1.13)	1138 (88.2)
Tprot	BMIxage; ethnicity	7.6	1.02 (0.98 to 1.07)	0.99 (0.96 to 1.03)	1.03 (0.99 to 1.07)	0.98 (0.93 to 1.02)	1159 (89.8)

All analyses use log-transformed analyte concentrations in the follow-up (FU1) panel.
ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; Bili, bilirubin; BMI, body mass index; GGT, γ-glutamyltransferase; Glob, globulin; Tprot, total protein.

Performance of alternative LFT panels

The numbers of TPs and FPs are plotted in figure 1 for each of the 255 possible LFT panels. The set of panels that are not dominated by any other panel is well approximated by the frontier in figure 1, defined by three panels involving just three analytes, ALP, ALP with ALT, and ALT with GGT. Of these, ALT and ALP arose as likely diagnostic candidates from the discriminant analyses, and GGT as the analyte with the highest overall positive rate (75.3%). The slopes of the line segments between the panels on the frontier (ie, ratio of extra TPs to extra FPs) are (ALP) to (ALP with ALT), 0.059 (SE 0.014); (ALP with ALT) to (ALT with GGT), 0.022 (SE 0.008). Thus, the single analyte panel (ALP) would be preferred so long as a TP is worth no more than the cost of approximately 17 FPs (=1/0.059); and the two-analyte panel (ALP with ALT), if this is more than 17 but less than 45 (=1/0.022). The analyte GGT (in combination with ALT) is not indicated unless the value of a TP is even higher. Furthermore, the slope of the line between (ALP with ALT) and the full panel is 0.021 (SE 0.007), suggesting that the full panel offers no enhancement unless the value of a TP is around 48 times the cost of an FP. The estimated PPVs for the panels on the frontier range from 8.7% (SE 2.8%) for ALP alone, through 6.3% (SE 1.2%) for ALP and ALT, to 4.4% (SE 0.7%) for ALT and GGT. The PPV of the eight analyte panel is 4% (SE 0.7%).

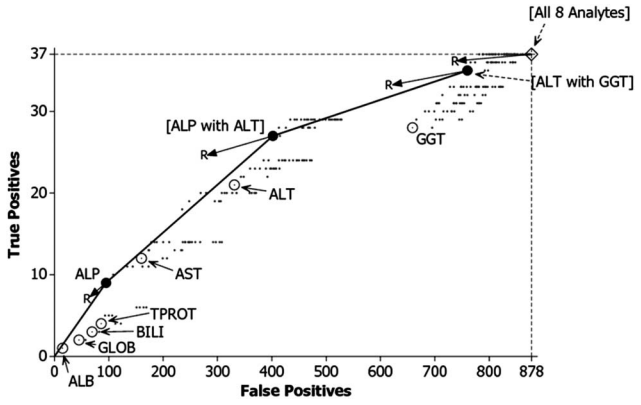

Figure 1 Positive diagnoses from different index panels. Split between the non-specific category (False Positives) and the pooled disease categories 1, 2 and 3 (True Positives). All 255 possible panels from the eight analytes are shown for the 915 patients with complete Index data. Single analyte panels (open circles) and the complete panel of eight analytes (diamond) are identified. The frontier (solid circles joined by line segments) shows the best diagnostic performance that can be attained using the analytes alkaline phosphatase (ALP), alanine aminotransferase (ALT) and γ-glutamyltransferase (GGT). The 2-analyte panel (ALP with aspartate aminotransferase (AST)) is also shown (open square). Results from repeating a panel at follow-up if it is positive initially are indicated by the letter 'R', joined by an arrow to the initial panel.

In the light of the results of the discriminant analysis, AST might be considered as an alternative to ALT in the construction of candidate panels. Indeed, the panels AST (PPV=7%, SE 2%) and (AST with ALP) (PPV=7.1%, SE 1.9%) generate similar PPVs to ALT and (ALT with ALP), respectively; but the overall yield, that is, total numbers of positives, is much reduced compared with the ALT versions. The panel (AST with ALP) generates only 257 positives compared with 429 for (ALT with ALP).

Repeat testing

The effect of repeat testing is also shown in figure 1. Here, it can be seen that repeating the full panel is an inefficient strategy achieving, for example, results similar to a single administration of a two-analyte panel (ALT with GGT; figure 1).

Effect of increasing the thresholds of abnormality

The full diagnostic value of individual analytes may not be captured by reference to conventional thresholds of abnormality. The effect of increasing these thresholds is investigated in figure 2 for the four analytes (GGT, ALT, AST, ALP) contributing the greatest numbers of positives in the LFT panel. For three of these (ALT, AST, ALP), the curves in the corresponding panels of figure 2 lie clearly above the diagonal line, showing that the ratio of TPs to FPs rises as the threshold increases. This entails an increase in PPV and is to be expected for markers that carry diagnostic information. For GGT, the ratio of TPs to FPs remains effectively constant as the threshold increases even to twice the conventional limit, rising only as it approaches a threefold increase. The effect of relaxing thresholds of abnormality cannot be determined given that entry to the study was based on conventional thresholds.

DISCUSSION

Principal findings

Our prospective study highlights the need to rethink the role of LFTs in primary care in the absence of obvious or pre-existing liver disease. First, the prevalence of significant liver disease in people with incidental abnormal LFTs in primary care is little higher than the general population prevalence (see below, *Meaning of the Study and Clinical Implications*). Second, repeating the full panel of LFTs is an inefficient strategy in primary care. Lastly, the results have potentially radical implications for the LFT panel in that selecting just two analytes (ALT and ALP) is an efficient strategy when the motivation for testing is the exclusion of significant liver disease: ALT is independently associated with specific hepatocellular diseases, while ALP is associated with biliary diseases and tumours.

Strengths and weaknesses of the study

BALLETS is a unique prospective study in that it comprised patients who presented in primary care with a

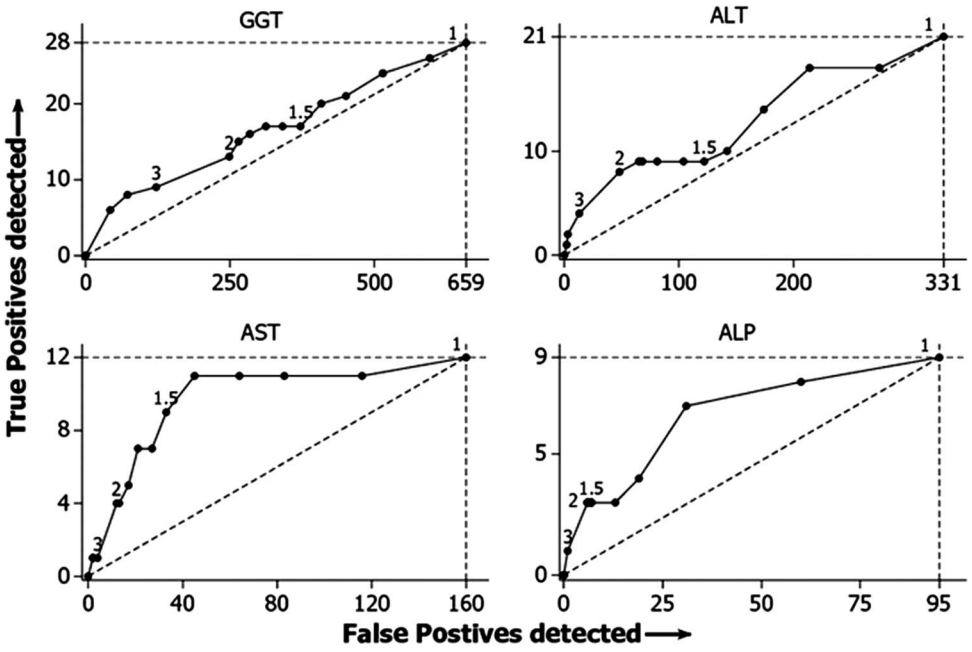

Figure 2 The effect of increasing the threshold of abnormality for four analytes. Numbers of True Positives (ie, patients in categories 1, 2 or 3 with analyte concentration above the threshold) are plotted against numbers of False Positives, using thresholds set at fixed multiples of the current laboratory reference limit. Points are plotted at intervals of 0.1 up to twice the reference limit, and at 3, 4 and 5 times the limit. Points at 1, 1.5, 2 and 3 times the limit are labelled accordingly.

history of liver disease, who were then comprehensively screened for liver disease and followed up for 2 years. The comprehensive screening 'compressed' future years, bringing forward diseases that might otherwise have presented only decades later. Documentation of clinical factors enabled analytes that were independently associated with various disease categories to be identified. In principle, this study can provide unbiased estimates of PPV of the LFT panel. However, when considered as a sample from a natural population, it is subject to selection bias since an abnormal Index LFT was a criterion for entry to the study. Consequently, attempts to measure the sensitivity or specificity of any particular combination of analytes will lead to biased estimates, despite the presence of normal analytes in the panel; sensitivities would be overestimated and specificities underestimated. Negative predictive value could not be measured. Evaluation of LFTs presents particular methodological challenges because, in contrast to the more usual one test/one disease scenario, up to eight analytes are involved and these may portend a large number of diseases. We dealt with the issue of many uncommon diseases by grouping them into clinically and pathologically meaningful categories, and the issue of multiple analytes by investigating the diagnostic capacity of all possible combinations leading to the 'frontier' in figure 1. Patients could only enter the BALLETS study when invited to do so by their GP. As part of the study, a sample of non-participating patients was compared with participants with respect to demographic features and severity of baseline abnormality. There was a small excess of older patients and patients with an abnormal GGT among those who participated in the study, but there was no difference in the degree of abnormality across these groups, suggesting that recruitment biases induced by GP behaviour are likely to be small.[2]

Strengths and weaknesses in relation to other studies

Previous primary care studies in the UK have been limited by their retrospective design,[3 6] in that identification of significant liver disease was dependent on investigations selected by the clinician and/or on review of hospital-based records. A recent record-linkage study[6] followed patients for a median of 3.7 years but without a full clinical investigation of the cohort. It reached similar conclusions with respect to the low overall predictive value of the LFT panel but ascribes greater importance to GGT than we have performed. However, the influence of selection effects on this conclusion cannot be discounted since GGT measurements were available for only 11% of the study sample.

The largest prospective data set outside the UK comes from the Dionysos study (n=6917), which was undertaken in two towns in northern Italy in the 1990s.[7] Although the study provided invaluable data on the prevalence of liver disease in a general (European) population, it could not extrapolate on the diagnostic performance of the full LFT panel currently being utilised by GPs, as only AST, ALT and GGT were collected. In contrast, our study provides unique information on individuals who are already engaging with local UK-health services (for a variety of health problems),

and thus the findings have instant ramifications for optimising the use of LFTs and preventing unnecessary investigations/repeat testing in primary care.

Meaning of the study and clinical implications

The PPV of the full LFT panel for specific disease affecting the liver/biliary tract is low in primary care. Less than 5% of people with an abnormal LFT panel had a specific liver disease and 1.7% needed specialised treatment (antiviral therapy or venesection for haemochromatosis). These findings are corroborated by the results of a recent record-linkage, which looked at PPV for mortality.[8] The prevalence of viral hepatitis and homozygous haemochromatosis (the most common categories of hepatocellular disease) in people with incidental abnormal LFTs was very close to population norms in England: 1% vs 0.7% and 0.5% vs 0.5%, respectively.[9–12] PPVs would most likely be higher in settings where liver disease, especially hepatitis B and C, is more common[13] (as in the Dionysos study). All but two cases of chronic viral hepatitis originated in moderate-risk to high-risk countries. If all such patients were screened (ie, HBsAg and HCV) when they first registered with a GP in the UK, the predictive values of LFTs would be lower still. We also considered a small number of diseases (discovered by reviewing case notes at 2-year follow-up) that could have affected LFT results (one case of Lyme disease, one of chronic pancreatitis and four cases of hypothyroidism). Including these six cases in the liver disease group leaves the findings essentially unchanged.[2] Figure 2 shows that the TP rate could be increased by raising the threshold of abnormality, but at the expense of the total number of TPs. We discuss the findings of the BALLETS study with respect to the enigmatic condition of fatty liver elsewhere.[2 14]

Selection of analytes

Our results suggest that the functions of a routine LFT panel can be largely subsumed into just two analytes: ALT and ALP. Furthermore, ALT and ALP contain information pointing towards definitive diagnosis, in that the former portends hepatocellular disease and the latter the biliary disease and tumours of the hepatobiliary system categories. In keeping with our results, Donnan *et al*'s[6] record linkage study highlighted that GGT had a high FP rate (figure 1). Our study casts further doubt on the clinical relevance of GGT by the finding that the PPV of an abnormal result (unlike those for ALT and ALP) does not demonstrate the expected increase when a higher threshold of abnormality is used (figure 2). Analytes apart from ALT and ALP should nevertheless be reserved for particular circumstances; for example, GGT and AST may be useful when it is suspected that a patient is in denial about alcohol intake, while Bili has a role when Gilbert's syndrome or acute hepatitis A is suspected.[2]

Repeat testing

There is a natural impulse to repeat a positive test to see if it is confirmed and this is the course of action recommended in current guidelines.[3 15–18] However, the results of the study show that this is an inefficient strategy when a full eight-analyte LFT panel is used in primary care when the sole purpose is to exclude significant liver disease in the absence of clinical signs. The impact of any test depends on events triggered downstream of the test itself.[19 20] The decision tree required to model the consequences of the full range of abnormal LFTs would be forbiddingly extensive and require untested assumptions such as the effect of various test results on unhealthy behaviours.[2] In a previous study, we modelled the cost-effectiveness of various strategies for the diagnosis of one serious treatable disease, chronic viral hepatitis, when the full index LFT is abnormal. It turns out that it is more efficient to test directly for the virus than to repeat the full liver panel with a view to viral testing if an abnormality persists. Performing a full panel LFT, with a view to repeating it if abnormal, was the least efficient option considered.[13] Although conducted with respect to a particular condition (viral hepatitis), this finding provides indirect support for the more general proposition that performing LFTs with a view to repeating them if abnormal is not the optimum strategy.

Acknowledgements We thank Peter Chilton (University of Birmingham) for help in preparation of this manuscript.

Contributors RJL was the lead applicant for the grant, chief investigator for the study, lead for the Birmingham site and wrote the manuscript. LMB project managed the study for the Birmingham site, performed the data collection and worked on the revision of all versions of the manuscript. MJA performed the clinical records review and worked on the revision of all versions of the manuscript. JN was a principal applicant on the grant and worked on the revision of all versions of the manuscript. AJG was the statistical applicant on the grant, did the data analysis and interpretation and cowrote the manuscript. RJL is the guarantor.

Funding The BALLETS study was financially supported by the National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme (HTA grant 03/38/01); AJG was financially supported by the Engineering and Physical Sciences Research Council (EPSRC) Multidisciplinary Assessment Technology Centre for Health (MATCH) programme (EPSRC grant number GR/S29874/01); RJL was financially supported by the National Institute of Health Research (NIHR) Collaborations for Leadership in Applied Health Research and Care (CLAHRC) for Birmingham and the Black Country; and MJA was financially supported by a Wellcome Trust Clinical Research Fellowship (grant number RCHX14302). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests None.

Ethics approval This study was conducted with the favourable opinion of the St Thomas' Hospital Research Ethics Committee and with the approval of the South Birmingham Primary Care Trust Consortium R&D Department.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Additional data are published in the full HTA report.

REFERENCES

1. Green RM, Flamm S. AGA technical review on the evaluation of liver chemistry tests. *Gastroenterology* 2002;123:1367–84.

2. Lilford RJ, Bentham L, Girling A, *et al*. *Birmingham and Lambeth Liver Evaluation Testing Strategies (BALLETS): a prospective cohort study*. Health Technology Assessment, vol. 17. 2013, pp 1–307.
3. Sherwood P, Lyburn I, Brown S, *et al*. How are abnormal results for liver function tests dealt with in primary care? Audit of yield and impact. *BMJ* 2001;322:276–8.
4. World Health Organisation. *Geneva: Good clinical laboratory practice (GCLP). World Health Organization on behalf of the special programme for research and training in tropical diseases*. Geneva: World Health Organisation Press, 2009.
5. StataCorp. *Stata statistical software: release 12*. College Station, TX: StataCorp LP, 2011.
6. Donnan PT, McLernon D, Dillon JF, *et al*. Development of a decision support tool for primary care management of patients with abnormal liver function tests without clinically apparent liver disease: a record-linkage population cohort study and decision analysis (ALFIE). *Health Technol Assess* 2009;13:1–156.
7. Bellentani S, Tiribelli C, Saccoccio G, *et al*. Prevalence of chronic liver disease in the general population of northern Italy: the Dionysos Study. *Hepatology* 1994;20:1442–9.
8. McLernon DJ, Dillon JF, Sullivan FM, *et al*. The utility of liver function tests for mortality prediction within one year in primary care using the Algorithm for Liver Function Investigations (ALFI). *PLoS ONE* 2012;7:e50965.
9. Health Protection Agency. *Hepatitis C Information*. Health Protection Agency, 2003. http://www.hpa.org.uk/infections/topics_az/hepatitis_c/phlsgen_info.htm (accessed 25 Mar 2013).
10. Health Protection Agency. *Hepatitis B information*. Health Protection Agency, 2003. http://www.hpa.org.uk/infections/topics_az/hepatitis_b/gen_info.htm (accessed 25 Mar 2013).
11. Worwood M. Haemochromatosis. *Clin Lab Haematol* 1998;20:65–75.
12. Olynyk JK, Cullen DJ, Aquilia S, *et al*. A population-based study of the clinical expression of the hemochromatosis gene. *N Engl J Med* 1999;341:718–24.
13. Arnold DT, Bentham LM, Jacob RP, *et al*. Should patients with abnormal liver function tests in primary care be tested for chronic viral hepatitis: cost minimisation analysis based on a comprehensively tested cohort? *BMC Fam Pract* 2011;12:9.
14. Armstrong MJ, Houlihan DD, Bentham L, *et al*. Presence and severity of non-alcoholic fatty liver disease in a large, prospective primary care cohort. *J Hepatol* 2012;56:234–40.
15. Muijtjens AM, Van Luijk SJ, Van Der Vleuten CP. ROC and loss function analysis in sequential testing. *Adv Health Sci Educ Theory Pract* 2006;11:5–17.
16. Pratt DS, Kaplan MM. Evaluation of abnormal liver-enzyme results in asymptomatic patients. *N Engl J Med* 2000;342:1266–71.
17. Giannini EG, Testa R, Savarino V. Liver enzyme alteration: a guide for clinicians. *CMAJ* 2005;172:367–9.
18. Theal RM, Scott K. Evaluating asymptomatic patients with abnormal liver function test results. *Am Fam Physician* 1996;53:2111–19.
19. De Bono M, Fawdry RD, Lilford RJ. Size of trials for evaluation of antenatal tests of fetal wellbeing in high risk pregnancy. *J Perinat Med* 1990;18:77–87.
20. di Ruffano L Ferrante, Hyde CJ, McCaffery KJ, *et al*. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ* 2012;344:e686.