A prospective single-centre evaluation of the intra-individual variability of faecal calprotectin in quiescent Crohn’s disease


*Department of Gastroenterology, Glasgow Royal Infirmary, Glasgow, UK. †Robertson Centre for Biostatistics, Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK. ‡Department of Clinical Biochemistry, Glasgow Royal Infirmary, Glasgow, UK.

Correspondence to: Dr Daniel R. Gaya, Gastroenterology Unit, Glasgow Royal Infirmary, G4 0SF, UK. E-mails: daniel.gaya@ggc.scot.nhs.uk; daniel.gaya@nhs.net

Publication data
Submitted 21 October 2012
First decision 16 November 2012
Resubmitted 2 January 2013
Accepted 3 January 2013

SUMMARY

Background
As a non-invasive marker of gastrointestinal inflammation, faecal calprotectin (FC) is being increasingly used to guide the management of Crohn’s disease. It is therefore a concern that studies have shown variability in day to day levels.

Aim
To determine the degree of this intrapersonal variability in the context of quiescent Crohn’s disease.

Methods
A single-centre prospective study was undertaken in 143 Crohn’s disease patients in clinical remission. Three faecal calprotectin levels were analysed from stool samples on consecutive days. Consistency of faecal calprotectin levels was determined by measuring the intraclass correlation (ICC). Due to higher variability at higher faecal calprotectin levels, the ICC was calculated for the log-transformed values. The reliability of detecting a ‘case’ of active inflammation as defined for specific concentrations of faecal calprotectin was measured by the kappa statistic.

Results
Ninety-eight complete sets of results were obtained. The ICC was 0.84 (95% CI: 0.79–0.89), which represents low variability across samples. The kappa statistic for the reliability of detecting a case as defined by an FC level of >50 μg/g was substantial at 0.648 (0.511–0.769).

Conclusions
Day to day variability of faecal calprotectin is low in our cohort of quiescent Crohn’s disease patients and the reliability of defining a ‘case’ is moderately good. These data provide reassurance to clinicians using a single calprotectin sample to inform therapeutic strategies in this cohort.
INTRODUCTION
The faecal calprotectin (FC) test has been shown to be useful as a non-invasive surrogate marker of the intestinal inflammatory burden in Crohn’s disease (CD) patients.1-3 It correlates well with faecal excretion of 111-indium-labelled leucocytes,4 the ‘gold standard’ marker of intestinal inflammation. Furthermore, it is resistant to protease breakdown meaning the protein is stable at room temperature in faecal samples for up to 7 days.5 As such it is used as a tool to aid CD diagnosis and management. FC has been shown to have a homogenous distribution in human faeces,5 supporting the notion that accurate analyses can be undertaken on a spot stool sample. It has been proposed that it may even be useful to predict mucosal healing at levels below threshold (50 μg/g),6 to potentially predict the risk of clinical relapse7-13 or to monitor response to biological therapies.14

However, these possibilities rely on the premise that levels of FC do not vary significantly from day to day in individual patients with CD. It is concerning that concentrations have been shown to be variable in those with active CD.1, 15 Any significant variability would call into question the practice of using a single sample to aid clinical decision-making. It is becoming routine to use the FC test to monitor response to inflammatory bowel disease (IBD) therapies. In addition, variable levels of FC (most commonly 50 and 100 μg/g) have been used to detect a ‘case’ of active inflammation16 and it would be useful to quantify how FC levels crossed these somewhat arbitrary boundaries in individuals. In the interests of quality assurance, we chose to investigate the level of variability within a cohort of CD patients in clinical remission, which has never previously been investigated in a suitably powered cohort. Our goal was to clarify the reliability and reproducibility of FC values for these individuals.

METHODS
Patients
In this single-centre prospective study, 143 consecutive CD patients in clinical remission attending for routine outpatient review between August 2010 and November 2011 were identified and enrolled. Written informed consent was obtained at the time of enrolment. Remission was defined as a Crohn’s disease Activity Index (CDAI)17 of <150 points.

Those aged over 18 years (no upper limit) with an existing definite diagnosis of CD in clinical remission as defined by a CDAI score of less than 150 were invited to participate.

We did not include patients with an unclear diagnosis (i.e. ‘indeterminate colitis’), clinical relapse within 3 months, chronic active disease requiring long-term steroid therapy, concomitant serious illness, pregnancy, age<18 years, alcohol abuse, nonsteroidal anti-inflammatory use and stool culture positivity.

Full ethical approval was awarded on 15 April 2010 by the West of Scotland Research Ethics Service (WeSRES) (REC reference 10/S0704/1).

Candidates with quiescent CD were invited to supply three stool samples for FC on three consecutive days. These samples were collected by the patients at home and processed at the biochemistry laboratory at Glasgow Royal Infirmary. Patients were encouraged to collect these at the weekend (Friday, Saturday & Sunday) and the three sample set for each patient was analysed within 48 hours of the final sample collection.

Biochemistry procedures
The stool samples were prepared and analysed according to the manufacturer’s instructions (Bühlmann calprotectin ELISA kit; BÜHLMANN Laboratories AG, Schönenbuch, Switzerland) using the Roche faecal extraction device. Stool was collected in screw-capped plastic containers and sent to the laboratory within 48 hours of the third stool collection. Processing was performed on site at Glasgow Royal Infirmary biochemistry laboratories by qualified biochemical scientists registered with the Health Professionals Council. Between 98 and 102 mg of faeces was weighed into extraction tube cap, 4.9 mL of extraction buffer was then added to all tubes which were recapped and homogenised for 15 min on the Alpha multi tube vortexer at maximum speed. The homogenate was centrifuged at 1849.8 g for 10 min and the supernatants transferred to plastic tubes and stored at −20°C. Time from sampling to preparation and freezing was estimated to be 1–3 days. The supernatants were thawed then mixed and centrifuged before analysis with the Bühlmann quantitative calprotectin ELISA kit on the Triturus automated ELISA analyser for determination of calprotectin in stools. Calprotectin was expressed as micrograms per gram (μg/g) of faeces. Faecal samples were stable between 2 and 8°C for up to 10 days and faecal extracts for 4 months at −20°C.

Statistical considerations
The reliability of the FC values across the three samples was measured by two methods, consistency and case reliability.
**Consistency.** The primary measure was the intraclass correlation (ICC) of the log-transformed FC values measured from the three samples obtained from each patient, measuring the proportion of the total variability (across all samples from all patients) that could be attributed to between-patient variation. This would be equal to its maximum value of 1 if there was no within-patient variation. Ninety-five per cent confidence intervals were calculated for the ICC.

**Case reliability.** The measure of case reliability was the kappa statistic for how well the three samples agreed in terms of whether the patient had a positive or negative result, where a positive result was defined as either FC > 50 μg/g or FC > 100 μg/g. This measure has a particularly useful clinical utility, providing clinicians with a guide as to the reliability of using FC to define whether or not a patient is a ‘case’ for a specific cut off.

The kappa statistic represents the agreement beyond chance. The value of the kappa statistic represents poor, slight, fair, moderate, substantial and almost perfect agreement when it is in the ranges of < 0, 0.00–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80 and >0.80 respectively. A kappa statistic of <0 represents agreement that is worse than chance. The 95% confidence interval for the kappa statistic was calculated using bootstrapping.

An exploratory analysis was also carried out to assess the case reliability when a case was defined by higher FC cut off values.

Both the consistency and case reliability analyses were repeated for patients with Crohn’s inflammation limited to the ileum.

**Sample size calculation**

The sample size calculation was based on a type I error of 0.05 and power of 80% for the primary measure. We estimated that the ICC would be 0.9, i.e. that 10% of the total variability would be due to within-patient variation between FC samples. For three replicates, a sample size of 95 patients would give a 95% confidence interval around the estimated ICC of (0.835–0.965). This sample size also gave 80% power with 5% significance for the case reliability measure (using the FC >50 μg/g cut off) to reject a null hypothesis of a kappa statistic of 0.5 (moderate agreement) in favour of the alternative that kappa was 0.9 (almost perfect agreement).

**RESULTS**

**Patients**

Of 143 patients recruited, 34 failed to return samples and six returned fewer than the required number. Two individuals experienced a flare of symptoms during the collection period and two withdrew consent. One patient’s sample leaked in transit. The analysis therefore included 98 patients with complete sets of results. Baseline characteristics for these patients are displayed in Table 1.

**FC values**

There were five patients with a total of eight FC values given by upper limits only (including six results of <30 μg/g and 2 of <10 μg/g), due to the lower limit of quantitation of the assay. These were assumed to have the upper limit value minus one (e.g. 29 for <30 μg/g) in the primary outcome analysis.

Although all patients were quiescent as defined by CDAI, the values showed wide-ranging and skewed distributions with most patients having relatively low values, while a few patients had very high FC values (Figure 1). Values ranged from close to 0 to around 2000 μg/g.

The three FC values for each patient are plotted against each other in Figure S1 (published online), which shows that there was a positive relationship between the

**Table 1 | Baseline characteristics of the 98 patients that took part in the study**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (mean ± s.d.)</td>
<td>47 ± 16</td>
</tr>
<tr>
<td>Gender (% male)</td>
<td>34%</td>
</tr>
<tr>
<td>Smoker (%)</td>
<td>19%</td>
</tr>
<tr>
<td>Stoma (%)</td>
<td>17%</td>
</tr>
<tr>
<td>CDAI (mean ± s.d.)</td>
<td>50 ± 39</td>
</tr>
<tr>
<td>CD location (%)</td>
<td></td>
</tr>
<tr>
<td>Ileal</td>
<td>15%</td>
</tr>
<tr>
<td>Ileocolonic</td>
<td>36%</td>
</tr>
<tr>
<td>Colonic</td>
<td>47%</td>
</tr>
<tr>
<td>CD phenotype (%)</td>
<td></td>
</tr>
<tr>
<td>Inflammatory</td>
<td>57%</td>
</tr>
<tr>
<td>Strictureng</td>
<td>30%</td>
</tr>
<tr>
<td>Fistulizing</td>
<td>12%</td>
</tr>
<tr>
<td>Perianal (%)</td>
<td>15%</td>
</tr>
<tr>
<td>5 ASA (%)*</td>
<td>43%</td>
</tr>
<tr>
<td>Corticosteroid (%)*</td>
<td>4%</td>
</tr>
<tr>
<td>Thiopurine (%)*</td>
<td>38%</td>
</tr>
<tr>
<td>Methotrexate (%)*</td>
<td>3%</td>
</tr>
<tr>
<td>Anti TNF (%)*</td>
<td>14%</td>
</tr>
<tr>
<td>CRP (mg/L) Mean ± s.d.</td>
<td>5.7 ± 7.6</td>
</tr>
</tbody>
</table>

* Medications refer to current use.
pairs of samples. There is an increase in variability between the values from any two samples as these values increase; showing a ‘fanning’ out as the values go from bottom left to top right in the plots. This is remedied by taking a log to base 10 (log10) transformation of the FC values, as displayed in Figure 2. The two rows show the same plots with different cut offs marked, which will be discussed in the case reliability section below.

The strong positive relationships between the FC values from each pair of samples are clear, with consistent variability across the range of log10 FC, suggesting that comparison of the FC values should be carried out on the log10 scale.

**Consistency.** The ICC was 0.84 (95% CI: 0.79–0.89) for the log FC values, showing that the overall consistency between the three samples from each patient in the study group is high. The narrow confidence interval suggests that agreement between FC samples would likely be similarly high in the wider patient population represented by this patient group. The ICC shows that a minority of about 16% of the variability between FC values in this study group was due to within-patient variability across their three samples. Therefore, based on this cohort, log-transformed FC may be considered a consistent measure.

**Case reliability.** Table 2 shows the classification as a ‘case’ or ‘normal’ of each patient according to their FC values from their three samples, where a case is defined as FC >50 μg/g or FC >100 μg/g. The total agreement across all three samples is above 70% for both cut offs, although FC >50 μg/g shows a higher level of consistency across the three samples than FC >100 μg/g.

Figure 2 also shows the extent of matching, with different shaped points to represent the mismatches. It appears that the sample 3 values tended to be slightly higher than the other sample values, so that patients were more likely to be classified as a case from sample 3, with this appearing to have a greater impact with the >100 μg/g cut off.

The overall reliability of detecting a case is given by the kappa statistic and its corresponding 95% confidence interval, for both cut offs, in Table 3. The kappa statistics and corresponding confidence intervals suggest moderate to substantial agreement in the study group for both cut offs. The FC >50 μg/g cut off shows marginally higher agreement than FC >100 μg/g.

The exploratory analysis of higher FC cut offs for case reliability showed that the best agreement across samples occurred when a case was defined as FC >350 μg/g. Table 2 shows that the total agreement was 86% and Table 3 suggests that in the wider patient population the case reliability using this cut off would likely be consistently substantial, if not in the almost perfect range.

**Ileal subset**

As an exploratory subset analysis, we carried out the consistency and case reliability analyses in those with Crohn’s inflammation limited to the ileum, owing to previous concerns on the reliability of the FC assay in this CD cohort.11, 12 There were 15 patients in the ileal subset. The ICC across the three log FC values for these patients was 0.816 (95% CI: 0.631–0.927), showing very slightly lower consistency than for the overall study group, but still high. However, the wide confidence interval shows that there is a great deal of uncertainty about this ICC value, due to the small number of patients in the subset.
The kappa statistic showing case reliability between the three samples was 0.59 (95% CI: 0.27–0.90) for FC >50 μg/g and 0.45 (95% CI: 0.11–0.73) for FC >100 μg/g. The agreement within the study group is therefore moderate for both cut offs, although on the verge of substantial for FC >50 μg/g. However, it could be anywhere from fair to almost perfect for ileal patients in the wider population. For the exploratory cut off of FC >350 μg/g the kappa statistic for the ileal subset was 0.88 (95% CI: 0.48–1.00), suggesting that agreement between samples could be anywhere from moderate to perfect agreement in the wider population.

**DISCUSSION**

Our dataset reveals that FC levels over a 3-day period in CD patients in remission are consistent and that one off spot stool values can indeed be utilised to aid clinical decision-making. Somewhat to our surprise, day to day variability was low within this cohort. Case reliability was moderate to substantial, particularly for a 50 μg/g cut off.

The utility of FC in the investigation and follow-up for luminal gastroenterology patients is ever expanding. Data support its use in differentiating irritable bowel syndrome (IBS) from inflammatory bowel disease (IBD),1, 22–25 evaluating abdominal discomfort,26 reducing the need for endoscopy in suspected IBD,27 assessing treatment response in IBD,14, 15, 28–30 predicting relapse in IBD,7, 8, 10–13 and predicting mucosal healing in IBD.3, 6, 30, 31 Indeed at our own institution, the monthly FC workload has increased from 50 samples a month in January 2007 to over 1000 per month in August 2012. With the prospect of rapid near-patient testing26 for immediate results at ‘one stop’ clinics its use may further increase.
Using FC as a predictive tool for clinical recurrence and for mucosal healing in CD requires analysis from clinically quiescent patients. The previous work done on FC variability has not specifically targeted this group. Two previous studies have raised the issue of intrapersonal variability of FC in active CD.\(^1\) \(^{15}\) Both studies show a significant variability in FC concentrations on day to day analysis. Tibble et al.\(^1\) studied 22 individuals with active CD (CDAI >150) over a 4-day period. The median coefficient of variation in calprotectin concentrations in the first morning stool over the 4 days was 54% (95% CI: 21–94) with the greatest variability at the lower concentrations. This level of variance is far greater than that of faecal elastase\(^3\) for example (15%). The level of FC concentration variance was higher than that for total FC excretion (29%). The authors postulate that stool volume and hydration may affect calprotectin concentrations, accounting for the variability increase. In our own study cohort, the coefficient of variation was calculated to be 20.9% (95% CI: 16.9–21.5) for log-transformed FC values. Moum et al.\(^{15}\) also studied an active CD population (\(n=68\)) with mild to moderate activity as defined by the Harvey Bradshaw Index (HBI). They chose to measure agreement rather than variance between two FC samples. Their data suggested a greater variability with a measure of agreement, kappa 0.355 (s.d.: 0.115; \(P<0.0002\)). They also chose to measure assay variability by analysing five subsamples from 10 stools and this variability was reassuringly low.

There is therefore a degree of variance within an active CD population, which may limit the utility of one off FC sampling in this cohort. Tibble et al\(^1\) state that the greatest variability was at low concentrations, which may mean a clinically quiescent population, would yield more variability. The good correlation across three samples shown in our study provides reassurance that sam-

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Percentage of the 98 patients classified as case or normal in each sample according to the definition of a case being FC &gt;50, &gt;100 or &gt;350 µg/g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case if FC &gt; 50 µg/g (total agreement = 79%)</td>
<td></td>
</tr>
<tr>
<td>3rd sample = Normal ((N=24))</td>
<td>2nd sample</td>
</tr>
<tr>
<td>1st sample</td>
<td>Normal</td>
</tr>
<tr>
<td>Normal</td>
<td>15%</td>
</tr>
<tr>
<td>Case</td>
<td>1%</td>
</tr>
</tbody>
</table>

Case if FC >100 µg/g (total agreement = 71%)

| 3rd sample = Normal (\(N=39\)) | 2nd sample | 3rd sample = Case (\(N=59\)) | 2nd sample |
| 1st sample | Normal | Case | Normal | Case |
| Normal | 34% | 1% | Normal | 5% | 5% |
| Case | 2% | 3% | Case | 13% | 37% |

Case if FC >350 µg/g (total agreement = 86%)

| 3rd sample = Normal (\(N=75\)) | 2nd sample | 3rd sample = Case (\(N=23\)) | 2nd sample |
| 1st sample | Normal | Case | Normal | Case |
| Normal | 69% | 1% | Normal | 5% | 0% |
| Case | 2% | 4% | Case | 2% | 16% |

The percentages agreeing across all three samples are in bold, while all nonbold percentages are mismatches. Any percentage discrepancies are due to rounding.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Reliability of defining a ‘case’ using calprotectin &gt;50, &gt;100 or &gt;350 µg/g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case definition</td>
<td>Kappa (95% CI)</td>
</tr>
<tr>
<td>Calprotectin &gt;50</td>
<td>0.648 (0.511–0.769)</td>
</tr>
<tr>
<td>Calprotectin &gt;100</td>
<td>0.603 (0.477–0.720)</td>
</tr>
<tr>
<td>Calprotectin &gt;350</td>
<td>0.732 (0.588–0.853)</td>
</tr>
</tbody>
</table>
pling this population will be reliable. We postulate that the variance may be reduced as asymptomatic patients will by definition have lower volume stools and that active disease in itself is more likely to produce inconsistent results due to changing inflammatory levels secondary to evolving disease or response to additional therapy.

The FC levels have also been shown to be variable in populations without IBD. Husebye et al.\textsuperscript{33} sampled stool from 14 patients referred for colonoscopy (without colonic inflammation or neoplasm) finding a coefficient of variability of 58%. Gilbert et al.\textsuperscript{34} studied 14 patients with known colorectal cancer finding a coefficient of variability for FC levels (mg/L) within individuals to be only 22%. But Husebye argues that these differences are mainly attributed to differences in mean estimates. If this is taken into account, the standard deviation within individuals is similar.

The test manufacturer states that an FC level $>50$ μg/g should define a positive case and studies comparing IBD cases and controls have used this value.\textsuperscript{23, 27, 30} In our study, both cut off levels (50 or 100 μg/g) showed moderately good reliability over the three samples with the 50 μg/g cut off proving slightly more consistent (Kappa 0.648 cf 0.603). However, other threshold values have been used\textsuperscript{1}–\textsuperscript{27} and the optimum value to determine remission in CD has not clearly been established. Published studies have used differing values to determine risk of relapse in this population.\textsuperscript{7}–\textsuperscript{13} A level of as high as $>340$ μg/g was used by Kallel et al.\textsuperscript{7} in colonic CD to define a risk of relapse with a good degree of accuracy (18.8-fold risk, sensitivity 80%, specificity 90.7%). The assay for FC used in this study differed from our own. However, we did determine that a cut off level of 350 μg/g was more reliable than either 50 μg/g or 100 μg/g to detect a 'case' – kappa 0.732 (0.588–0.853). Our own results show that when a cut off of 100 μg/g was used to predict relapse, 29% of patients had results which fell either side of this marker over a 3-day period, whereas when a cut off of 350 μg/g was used, this figure dropped to 14%. These data provide reassurance that 350 μg/g is comparatively stable in terms of variability if it proves useful as a candidate cut off level for predicting relapse or monitoring treatment response. This is interesting as, with increasing clinical experience of FC in CD, it has become apparent that the ‘normal’ reference range of a FC in CD patients in remission is often far higher than the manufacturer’s guidance range; it may be that the latter would be more appropriately used in the general, as opposed to the CD, population.

Studies investigating CD relapse rates have reported that FC concentration may be a less reliable predictor of relapse in ileal CD.\textsuperscript{11, 12} We included 15 patients of this subtype in our analysis. When analysed separately we discovered that both consistency and case reliability were lower for these patients than for the whole study group. Confidence intervals were wide, suggesting a larger group of ileal patients would be necessary to come to a conclusive answer regarding reliability of FC in this population.

There are weaknesses in our study which merit consideration. No subsample analysis was performed for validation purposes. Although we are reassured that Moum et al.\textsuperscript{15} found low variability, this would have served to quality assure our own ELISA method for this population. We also did not confirm whether our patients were in endoscopic remission at the time of enrolment. Although this approach exposed the candidates to less risk and inconvenience, we accept that we have not provided ‘gold standard’ evidence of remission. We encouraged candidates to predominantly collect their samples at the weekend as this was thought to be more convenient for individuals and to allow more time for laboratory processing. It is possible, however, that this timing could affect variability levels due to dietary differences for individuals at weekends compared with weekdays. The study was conducted within a tertiary referral centre with a motivated patient population that may be skewed towards deep remission. This could make the data less applicable to the wider population and specifically those who have recently entered remission through induction therapy.

In summary, these data reveal, for the first time, that the day to day variability in the FC test result is low in a large cohort of patients with CD in remission. The reliability of the standard cut off levels is also moderately good. This provides reassurance for healthcare workers seeking to utilise it as a predictive test and a tool to aid decision-making in this population.

AUTHORSHIP
Guarantor of the article: D. R. Gaya.
Author contributions: Dr Naismith and Dr Gaya designed the study. Dr Rankin coordinated processing of samples. Ms Munro, Ms Laird, Dr Morris, Dr Winter, Dr Gay and Dr Smith collected the data. Dr Smith collated the data. Dr Barry performed statistical analysis and wrote the statistics section of the paper. Dr Naismith wrote the paper with contributions from Dr Smith, Dr Gaya and Dr Rankin. All authors approved the final version of the manuscript.
ACKNOWLEDGEMENTS

Declaration of personal interests: None.

Declaration of funding interests: The biochemistry processing and statistical analysis required financial support which was provided by an educational grant from ‘Cure Crohn’s colitis’ (C3) http://www.curecrohnscolitis.org - charity number SC036559.

REFERENCES

29. Palmon R, Brown S, Ullman TA, et al. Calprotectin and lactoferrin decrease with...
maintenance infliximab administration
[abstract]. Gastroenterology 2006; 130
(Suppl. 2): A-212.
30. Langhorst J, Elsenbruch S, Koelzer J,
Rueffer A, Michalsen A, Dobos GJ.
Noninvasive markers in the assessment
of intestinal inflammation in
inflammatory bowel disease:
performance of fecal lactoferrin,
calprotectin and PMN-elastase, CRP
and clinical indices. Am J Gastroenterol
31. Jones J, Loftus EV, Panaccione R,
et al. Relationships between disease
activity and serum and fecal
biomarkers in patients with Crohn’s
disease. Clin Gastroenterol Hepatol
32. Loser C, Mollgaard A, Folsch UR.
Faecal elastase: a novel, highly sensitive,
and specific tubeless pancreatic function
33. Husebye E, Ton H, Johne B. Biological
variability of fecal calprotectin in
patients referred for colonoscopy
without colonic inflammation or
neoplasm. Am J Gastroenterol 2001; 96:
2683–7.
34. Gilbert JA, Ahlquist DA, Mahoney
DW. Fecal marker variability in
colorectal cancer: Calprotectin versus
haemoglobin. Scand J Gastroenterol