Glomerular C3 staining is an independent predictor of renal survival and eGFR slope in IgA nephropathy: findings from the UK RaDaR cohort and its nested GPT-4o data extraction study

Sherry Masoud[1,2], David Pitcher[1,2], Katie Wong[1,2], Yunsoo Kim[2], Alex Shavick[3], Adam P Levine[2], Jonathan Barratt[4], Daniel P Gale[1,2], Ian SD Roberts[5]

[1] UK National Registry of Rare Disease (RaDaR), Bristol, UK [2] University College London, UK [3] Human Technopole, Milan, Italy [4] University of Leicester, UK [5] University of Oxford, UK.

Introduction

IgA nephropathy (IgAN) is the most common primary glomerulonephritis worldwide. Activation of the alternative complement pathway plays a key role in IgAN pathogenesis. The prognostic value of glomerular C3 deposition remains uncertain, particularly outside East Asian populations. Clarifying this relationship could help define the role of kidney biopsy in guiding complement-targeted therapy and facilitate precision medicine within an evolving therapeutic landscape. Renal pathology data are rich in prognostic information but largely unstructured, making real-world data analyses labour-intensive and difficult to scale. Accurate large-scale data extraction would address a key bottleneck, facilitating the effective use of pathology data in research and potentially support future image analysis applications. The RaDaR registry provides a unique opportunity to investigate the prognostic value of glomerular C3 in IgAN and to test whether a GPT-4o based pipeline can reliably extract pathology data at scale.

Objectives

1. Evaluate the association between glomerular C3 deposition (assessed by immunofluorescence) and renal outcomes (eGFR slope and renal survival) using manually extracted pathology data.
2. Assess the performance of GPT-4o (via an Azure privacy-preserving platform) for semi-automated extraction of pathology features, using the manually curated dataset as ground truth.
3. Compare clinical outcome associations derived from manual versus GPT-4o extracted datasets.

Methods

Design: Retrospective cohort study using data from the RaDaR Registry (UK Registry of Rare Renal Diseases)
Inclusion criteria: age ≥16 years; index biopsy after 2010; primary IgA nephropathy (excl. HSP).
Exclusion criteria: <8 glomeruli; eGFR <30 mL/min/1.73 m² or kidney failure; <6 m follow-up.
Analytic Subsets: 571 biopsies; 100 randomly sampled for prompt development; remaining 471 formed an independent cohort for performance & outcome analyses.
Outcomes: Renal survival (defined as kidney failure or death) & eGFR slope.

Data Extraction Methods
Ground truth: Two specialists extracted data. C3 scored 0–3 and dichotomized <2 vs ≥2 for analyses.
Prompt development (n=100): Human-in-the-loop, error-driven prompt refinement (schema/offset checks, rule-based constraints); prompt frozen for the 471-case run.
GPT-4o extraction (Azure OpenAI, 2024-05-13): Constrained-JSON output returned verbatim evidence spans and provisional C3/MEST-C scores; a deterministic rules layer verified spans and applied pre-specified criteria, abstaining if evidence was insufficient.

Statistical Analysis and GPT-4o Performance Metrics:
- Ordinal logistic, Cox PH & linear mixed-effects models were used to examine association of C3 Staining with 1) MEST-C scores 2) Renal Survival 3) eGFR slope (co-variates age, sex, eGFR , urine PCR & MEST-C).
- Performance metrics included coverage, abstain rate, quadratic-weighted κ, precision, recall, F1 score, and specificity. Reproducibility was assessed across three runs using the paired Wilcoxon SR test.
- Agreement between manual and GPT-4o derived C3 effects in Cox models was assessed with 1,000 paired bootstrap resamples of the difference in β (log-HR).

Results

There were **471 patients** included in the evaluation cohort. The mean age was **43 years** and 67% were male. Median (IQR) eGFR was **57mL/min/173m²** (42-82) and PCR **165mg/mmol** (63-327). Median (IQR) **follow-up** was **8.4** years (3.2-9.5). 144/471 **(31%)** experienced **kidney failure** or **death** as **first event**. **272 (61%) biopsies** had **C3 intensity** reported as **≥ 2**. Baseline characteristics by C3 intensity can be found in Table 1 (excluding those with non evaluable C3 results). C3 intensity ≥2 was associated with ~**1.5×** higher adj. **odds** of **elevated M,S and T scores** (Table 2).

Cox PH showed high C3 intensity (**HR 2.36, 1.28–4.35**), lower baseline eGFR (per 5 mL/min: **HR 0.83, 0.77–0.90**), and higher baseline PCR (per doubling: **HR 1.30, 1.07–1.57**) were independently associated with **kidney failure or death**. (Table 3)

Table 1 . Baseline characteristics by C3 Intensity

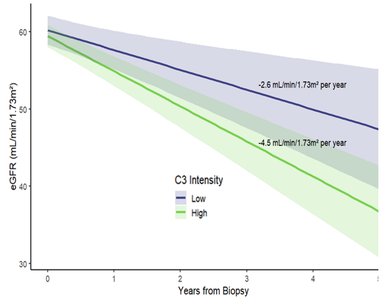
	C3 Intensity <2 N=177	C3 Intensity ≥ 2 N=272
Male	115 (65%)	190 (70%)
Age (SD)	45 (13.9)	42.1 (15.8)
Caucasian	119/150 (79%)	204/253 (81%)
Follow-up years (IQR)	8.1 (4.2 - 9.1)	8.6 (2.8 - 9.8)
Baseline eGFR (IQR)	56 (42 - 82)	57 (42 - 82)
Baseline PCR mg/mmol (IQR)	182 (73 - 308)	159 (59 - 372)
ARB/ACEi where data available	33/46 (72%)	49/78 (63%)
M= 1	89 (51%)	162 (60%)
E= 1	47 (27%)	84 (31%)
S= 1	116 (66%)	204 (76%)
T≥ 1	59 (33%)	118 (44%)
C≥ 1	58 (33%)	70 (26%)
Kidney failure	43 (24%)	86 (32%)
Kidney failure/death as 1st event	45 (25%)	92 (34%)

Table 2. Is C3 associated with higher M,E,S,T and C scores ?

	OR*	95% CI	P value
M1 score	1.5	1.05–2.13	**0.026***
E1 score	1.28	0.87–1.91	0.21
S1 score	1.58	1.08–2.32	**0.019***
↑T score	1.59	1.11–2.28	**0.011***
↑C score	0.79	0.54–1.16	0.236

*adjusted for age/sex

Figure 1. 5-year eGFR trajectory stratified by C3 intensity on biopsy (Linear mixed model)*



* Adjusted for PCR, eGFR, Age, Sex, MEST-C scores

Table 2. Clinical & histological risk factors for kidney failure or death as first event

	Univariable Analysis			Multivariable Analysis		
	HR	95% CI	P value	HR	95% CI	P value
Age (per 10 year)	1.19	1.07-1.32	0.001*	0.85	0.68-1.05	0.122
Female	0.58	0.40-0.85	0.006*	0.89	0.51-1.58	0.697
Baseline eGFR (per ↑5 mL/min/ 1.73m²)	0.85	0.81-0.90	<0.001*	0.83	0.77-0.90	**<0.001***
Baseline PCR† (per doubling)	1.19	1.05-1.34	0.005*	1.30	1.07-1.57	**0.007***
Histological Findings						
M1	1.14	0.81-1.58	0.456	0.72	0.41-1.27	0.257
E1	1.38	0.97-1.95	0.073	1.3	0.70-2.40	0.408
S1	2.01	1.33-3.06	0.001*	1.33	0.59-3.00	0.486
T ≥ 1	3.07	2.21-4.27	<0.001*	1.21	0.68-2.14	0.513
C ≥ 1	1.01	0.71-1.45	0.947	0.74	0.38-1.44	0.383
C3 Staining ≥ 2	1.38	0.98-1.97	0.069	2.36	1.28-4.35	**0.006***

†Baseline PCR (mg/mmol) was log-transformed in the Cox model; HRs & 95% CIs are expressed per doubling of baseline PCR for interpretability

Table 4. GPT-4o extraction of C3 and MEST-C vs. manual ground truth

Performance of GPT4o at extracting C3 scores						
	Coverage	Recall* (Sensitivity)	Precision* (PPV)	F1 Score *	Specificity*†	Quadratic κ
C3	97.1%	0.899	0.896	0.896	0.970	0.919
Performance of GPT4o at inferring MESTC scores						
	Coverage	Recall‡ (Sensitivity)	Precision‡ (PPV)	F1 Score ‡	Specificity†‡	
M	98.0%	0.851	0.896	0.873	0.836	
E	88.5%	0.837	0.954	0.891	0.984	
S	81.5%	0.974	0.982	0.978	0.936	
T	98.5%	0.822	0.827	0.811	0.850	
C	91.0%	0.934	0.932	0.933	0.953	

*weighted averages † Specificity is computed one-vs-rest per class if the variable is not binary. ‡ Weighted averages for T and C scores

Results continued

Adjusted **5-year eGFR slope** showed a **steeper decline** in patients with **C3 intensity ≥2 (–4.5 vs –2.6** mL/min/1.73 m²; p=0.038; Figure 1).

GPT-4o yielded **valid C3 outputs** for **436 of 449 scorable cases** (**97% coverage**) and abstained on 34 of 471 total requests (7%). Among 22 non-evaluable cases, 19 abstentions (86%) were appropriate. **C3 extraction** achieved high performance, with **precision**, **recall**, and **F1 scores** near **0.90** and **specificity 0.97** (Table 4). **Quadratic κ = 0.919**, reproducible across two additional runs (0.919 and 0.917; p = 0.50).

As reported above, **C3** was associated with **kidney failure or death (HR 2.36)**. Across **1,000** paired bootstrap **resamples**, the **C3 effect** estimated from LLM-extracted data **differed from manual** abstraction by **only 6%** (**HR 0.94**; 0.65–1.27) and was **not statistically significant** (p = 0.67).

GPT-4o also successfully **inferred MEST-C scores** from free-text reports when **not explicitly documented** (Table 4).

Conclusion

- **High glomerular C3 deposition was associated with more severe histological injury, faster eGFR decline, and an increased risk of kidney failure or death, supporting its role as a prognostic marker in IgAN.**
- **GPT-4o accurately and reproducibly extracted key pathology features from unstructured biopsy reports with performance comparable to clinician-curated data, substantially reducing manual effort.**
- **Clinical outcome associations derived from LLM-extracted data closely matched those from manual abstraction, demonstrating the feasibility of scalable semi-automated pathology data extraction.**

Contact Information : Sherry Masoud, sherry.masoud@ukkidney.org